# Trend Estimation on Social Media

## D1.7: Intermediary social media analysis report & visualisations

| | |
|---|---|
| **Work package** | WP 1: Topic Identification |
| **Task** | 1.7 NGI Topics Filtering & deep dives |
| **Due date** | 31/12/2019 |
| **Submission date** | 29/12/2019 |
| **Deliverable lead** | DATALAB, Aarhus University |
| **Dissemination level** | Public |
| **Nature** | Report |
| **Authors** | Lynge Asbjørn Møller, DATALAB<br>Kristoffer Laigaard Nielbo, CHCAA & DATALAB<br>Peter Bjerregaard Vahlstrup, CHCAA & DATALAB<br>Anja Bechmann, DATALAB |
| **Version** | 1.0 |
| **Reviewers** | Heshani Jayaratne, Nesta<br>Katja Bego, Nesta<br>Kristóf Gyódi, DELab UW<br>Michał Paliński, DELab UW<br>Łukasz Nawaro, DELab UW |
| **Status** | Ready for submission |

**Disclaimer**: The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained herein.

**Acknowledgement**: This Report is part of a project that has received funding from the **European Union's Horizon 2020 research and innovation programme under grant agreement Nº825652**

# Executive Summary

This report details the development of a new model for trend estimation and the intermediary results from applying the model on the American social news, web content and discussion website Reddit.com with the aim of delivering insights into the existing and emerging online social media discourse on internet-related topics.

Accurate trend estimation is a matter of debate in the research community, and the standard approaches often suffer from several methodological issues by focusing solely on spiky behaviour and thus equating trend detection with that of natural catastrophes and epidemics.

These problematic issues are remedied by domain knowledge of social media combined with advances in information theory and dynamical system in a new approach to trend estimation in which trend reservoirs – signals that display trend potential – are identified by their relationship between novel and resonant behaviour, and their minimal persistence.

The model estimates Novelty as a reliable difference from the past – how much does the content diverge from the past – and Resonance as the degree to which future information conforms to the Novelty – to what degree does the novel content 'stick'. Using calculations of Novelty and Resonance, trends are then characterized by a strong Novelty-Resonance association and a higher Hurst parameter in comparison to a random baseline.

Results show that these 'signatures' capture different properties of trend reservoirs, information stickiness and multi-scale correlations respectively, and they both have discriminatory power, i.e. they can actually detect trend reservoirs.

In this report, the new model for trend estimation is applied for deep dives into Reddit discussions related to artificial intelligence to exemplify the application of the model. The most trending subreddits are discovered using the model on a sample of subreddits with the highest overlap between their descriptions and a seed list of AI-related terms. After the classification, the content on the most trending subreddits is explored by training a neural embedding model to query the highest-ranking words and their associated words, provided insights into e.g. the contexts in which certain technologies are discussed.

This trend estimation model can be developed into a recommender system by using the classifier to train a recommender engine to identify trending subreddits within any given subject. Classifications of trending subreddits can be extremely useful for decision support in terms of which subreddits to follow for a continuum of information on trends topics such as AI.

# Table of Content

# 1 Introduction

Sociocultural trends from social media platforms have become an important part of knowledge discovery. The growth of social media usage over the last decade has opened up a data trove for researchers to analyse patterns in communication, allowing them to gain insights into new trends or emerging issues (Stieglitz, Mirbabaie, Ross, & Neuberger, 2018, p. 156).

The ‘trend’ construct is however ambiguous and its estimation from unstructured sociocultural data is complicated by several methodological issues. In this report, a new approach to trend estimation is presented and used to discover trends in social media discussions on internet development. The new approach was first proposed by Nielbo, Vahlstrup, Gao, & Bechmann (2019) and combines (‘intersects’) domain knowledge of social media with advances in information theory and dynamical systems. In particular, trend reservoirs (i.e., signals that display trend potential) are identified by their relationship between novel and resonant behaviour, and their minimal persistence.

## 1.1 Purpose and scope

The new model for trend estimation was developed with the aim of delivering insights into the existing and emerging online social media discourse on internet-related topics. This report details the intermediary results.

Social media can provide us with rich insights about which internet-related topics are frequently discussed and which topics are emerging in the discussions, allowing us to identify key issues and technologies and the ways in which the online discussions are shifting. The insights from social media will help inform outreach and dialogue with diverse stakeholders shaping the future internet and inform the future EU research agenda and funding as part of the *Next Generation Internet* initiative.

The focus of this report is internet-related discussions on the social media site Reddit. Reddit hosts discussions about text posts and web links across hundreds of topic-based communities called “subreddits” that often target specialized expert audiences on these topics (Horne, Adali, & Sikdar, 2017, p. 1). Topically defined discussions are thus an important part of the appeal of Reddit, unlike the information dissemination focus of Twitter (Kwak, Lee, Park, & Moon, 2010), and the specialized audiences make it a promising source for topical discussions on i.e. internet technology.

Applying the new model for trend estimation to Reddit discussions, we can identify trending subreddits related to internet development and analyse how and around which topics the discussions within these subreddits evolve and develop.

In this report, the application of the new model for trend estimation is exemplified through deep dives into the most trending subreddits related to *artificial intelligence*. The trending subreddits are identified by applying the model on a sample consisting of the top 100 subreddits with the most overlap in subreddit descriptions and a list of trending AI-terms identified by Gyódi, Nawaro, Paliński, & Wilamowski (2019) in *D1.2: Visualisations of key emerging technologies and social issues.*

The most trending of these subreddits are then further investigated through deep dives into the discussions. We train a neural embedding model to query the highest-ranking words and their associated words and build content graphs of these word contexts, allowing us to study the contexts in which technologies are being discussed.

The intermediary results featured in this report will be expanded upon in *D1.8 Final social media analysis report & visualisations* to be delivered in 2021.

The future report will also feature an analysis of Twitter discussions revolving around internet-related hashtags, identifying the networks, connections and communities surrounding these topics. As Twitter does not allow for the collection of historic data through its public APIs, the data collection is ongoing and the empirical basis for analysis is thus not sufficient enough for a meaningful analysis at this point of time. While historical Twitter data is available through Twitter's fully archived search API for a fee, these costs are not within the budget of the project. Other options include open data sources such TweetsKB, but they are most often not updated regularly (TweetsKB have not been updated since March 2018), which does not align with the aim of delivering insights into the current and emerging online social media discourse on internet-related topics. Instead, we are currently collecting all tweets with internet-related hashtags through Twitter's public Streaming API.

# 2 Methods

This section describes data, models and analysis for estimation of trend reservoirs and content exploration on the American social news, web content and discussion website Reddit.com. See Appendix A for more details and the equations behind the model for trend estimation.

## 2.1 Data and samples

The study uses all post titles from two samples (trending vs. not trending) of subreddits from Reddit.com. Subreddits are niche fora that discuss topics related to a forum subject (e.g., *r/MachineLearning*) and titles represent a uniform and comparable data element across all subreddits - titles rely only on natural language and are hosted at Reddit.com even when they refer to external websites. NSFW subreddits, which are typically of explicit pornographic content, that contain sample relevant terminology (e.g., *deepfake*) have been excluded from the samples. In the top 100 of the target AI samples they only occur three times.

### 2.1.1 Design and Statistical analysis

Sampling is based on content overlap in subreddit descriptions and an expert-based seed list delivered as part of *D1.2: Visualisations of key emerging technologies and social issues*. For the Artificial Intelligence sample the seed list was {'ai', 'facial recognition', 'project maven', 'reinforcement learning', 'ai startup', 'ai ethic', 'neural network', 'ai strateg*','machine-learn', 'ai algorithm', 'ai research', 'deep fake'} (Gyódi et al., 2019).

The trending sample consisted of the subreddits with the greatest word overlap in their description (Community Details and Rules) for each set (e.g., *r/artificial* and *r/MachineLearning* for *AI*) with the constraint of minimum 256 posts for accurate estimates of the Hurst parameter. While it is possible to get accurate estimates with as little as 120 data points (Gao, Hu, Mao, & Perc, 2012), larger samples will ensure robustness across methods.

The not trending sample was a random sample (without replacement) of subreddits with the constraints that the subreddits were not in the trending sample and they had at least 256 posts.

## 2.2 Model for estimation of trend reservoirs

In the typical case of trend estimation for social media, a query term (e.g., ‘AI’) is used to extract a signal based on the term’s frequency, associated queries, and rating systems. While researchers agree that a trend has direction (e.g., an increase in AI-related posts) and tendency (e.g., “AI is the new black”), accurate estimation is a matter of debate (Gray, 2007).

In its simplest form, a trend’s tendency is detected as a ‘Novelty spike’ in the query’s temporal distribution and the direction is estimated as the slope coefficient of the query’s frequency fitted on time (e.g. Madani, Boussaid, & Zegour, 2014; Mathioudakis & Koudas, 2010). This *standard approach* suffers from several problematic issues: 1) by focusing on spiky behaviour, it equates a sociocultural trend detection with that of natural catastrophes and epidemics; 2) it makes strong assumptions on the trend's shape; 3) it treats atomic words as semantically meaningful; and in pre-selecting query terms it 4) can fail to establish a proper baseline; and 5) reverse time order by nominating queries that show a spiky behaviour in the past as future trends.

These five issues can be remedied by techniques from information theory and dynamical systems theory. Recent studies have shown that windowed relative entropy can generate signals that capture information Novelty as a reliable difference from the past and Resonance as the degree to which future information conforms to the Novelty (Barron, Huang, Spang, & DeDeo, 2018; Murdock, Allen, & DeDeo, 2017; Kristoffer L. Nielbo, Perner, Larsen, Nielsen, & Laursen, 2019). Several studies have used latent semantic models to summarize the data set's co-occurrence structure as an alternative to atomistic query terms (Chinnov, Kerschke, Meske, Stieglitz, & Trautmann, 2015; Stieglitz et al., 2018). Regarding the trend shape, a smoothing function that fits piecewise polynomials to the data makes no assumption about the shape (Gray, 2007; Tenen, 2018). Recently, dynamical systems approaches have indicated that adaptive functions hold great promise for smoothing sociocultural data (Gao, Hu, Mao, & Perc, 2012; Gao, Jockers, Laudun, & Tangherlini, 2016; Hu, Liu, Thomsen, Gao, & Nielbo, 2019; Nielbo, Baunvig, Liu, & Gao, 2019; Wevers, Gao, & Nielbo, 2019)

In this report, these insights are combined in a new approach to trend estimation first proposed by Nielbo, Vahlstrup, Gao, & Bechmann (2019). This model studies 'trend reservoirs' which are characterized by a strong Novelty-Resonance association and a higher Hurst parameter in comparison to a random baseline. The approach to the estimation of trend reservoirs generalizes to other data sources (e.g., Twitter) and types (e.g., images). For stable estimates, a signal has to consist of a minimum of 256 data points (e.g., posts in a subreddit).

### 2.2.1 Topic modelling

To estimate a trend, we need a structured semantic representation of the data (in this case Reddit titles). This can be obtained through a simple and computationally efficient technique with Latent Dirichlet Allocation (LDA), which trains a probabilistic model on the data in order to create dense low-rank vector representations. LDA model is a popular method for implementing *topic modelling*, that can be used to extract topic information from large collections of texts (Blei, 2012). A topic model represents each text (in this case each Reddit title) as a mixture of latent topics based on the text's word co-occurrence structure.

### 2.2.2 Novelty, Transience and Resonance

Now that we have a structured semantic representation of our data, we can begin to estimate the trend potential of the data by estimating *Novelty*, *Transience*, and *Resonance* for a fixed time window of three days (see equations in Appendix A).

Novelty captures how much the content diverges from previous content. Looking at Reddit data, Novelty is the degree to which the subreddit produces discussions that diverge from the previous discussions. High Novelty signifies that many new topics are being introduced in the subreddit.

Similarly, Transience captures the degree to which the content differs from future content. In the context of our investigations of Reddit that means the degree to which newly introduced topics fade away in future discussions in the subreddit. High Transience signifies very shifting discussions.

Finally, Resonance is the difference between Novelty and Transience - the quality of at once differing from the past and leaving traces on the future. In the context of Reddit, posts with high Novelty and low Transience, and thus high Resonance, introduce novel content that 'sticks' so to say, and successfully changes the future of the discussions.

With the calculations for Novelty, Transience, and Resonance, we can then estimate the trend potential, an indicator variable for trend reservoirs, as the linear slope coefficient ($\mathbb{N} * \mathbb{R}$) (see equations in Appendix A). On Reddit, a subreddit's trend potential is estimated as its posts' *Resonance* on *Novelty*. Comparing the trend potential of human annotated 'trending' subreddits with randomly selected subreddits, Nielbo, Vahlstrup, Gao, & Bechmann (2019) showed that trend reservoirs have steeper ($\mathbb{N} * \mathbb{R}$) slope in comparison to the random baseline. Results indicate that $\mathbb{N} * \mathbb{R} > 0.77$ is a signature of trend reservoirs.

Below, the classification of a subreddit as trending or not trending based on the linear slope coefficient ($\mathbb{N} * \mathbb{R}$) of its posts' Resonance on Novelty is exemplified with the subreddit *r/MachineLearning* (Figure 1).

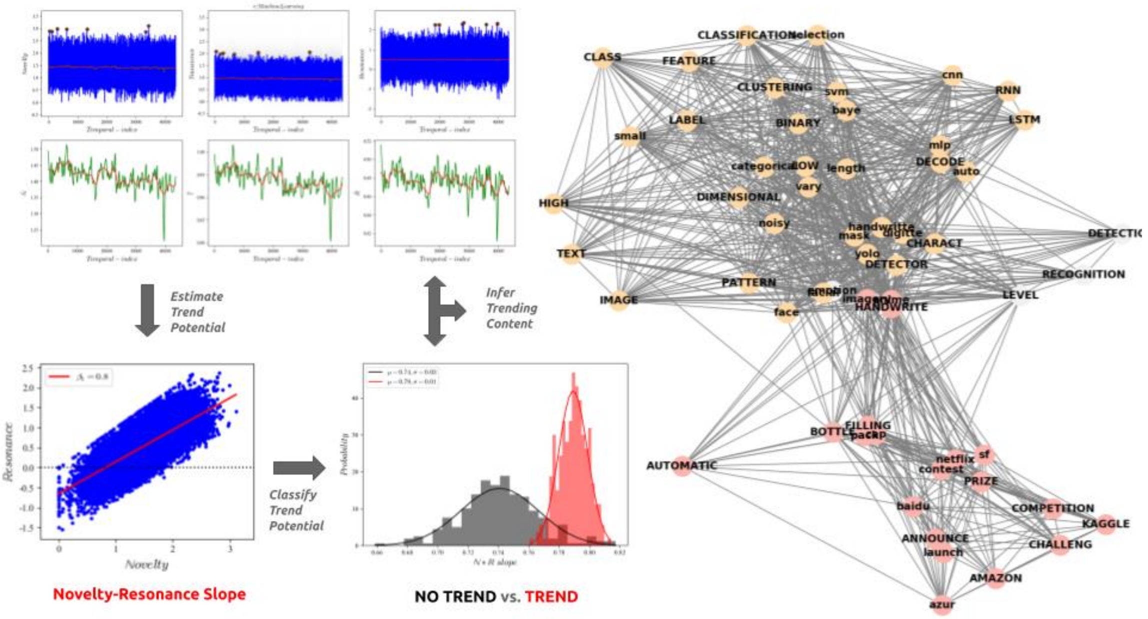

Figure 1: *Identification of trend potential as the linear slope coefficient of its posts' Resonance on Novelty on streaming data from r/MachineLearning on Reddit (find all figures in Appendix B)*

We start by computing information Novelty, Transience and Resonance. In Figure 1, Novelty, Transience and Resonance for *r/MachineLearning* are illustrated with adaptive filtering in the upper left corner (Figure 1a in Appendix B). The subreddit's trend potential is then estimated as the linear slope coefficient $(\mathbb{N} * \mathbb{R})$ of its posts' Resonance on Novelty as described above (the Novelty-Resonance Slope in the lower left corner of Figure 1, see also Figure 1b in Appendix B).

Based on the subreddit's trend potential, we can then classify the subreddit as either trending (red) or not trending (black) on Reddit (NO TREND vs. TREND in Figure 1, see also Figure 1c in Appendix B). With a linear slope coefficient of $\mathbb{N} * \mathbb{R} = 0.8$ , the *r/MachineLearning* subreddit can be classified as trending based on its posts' Resonance on Novelty. Finally, we infer the associated content on trending subreddits (graph on the right of Figure 1, see more in section 2.3).

### 2.2.3 Long-range memory

Other than a strong Novelty-Resonance association, trend reservoirs are also characterized by a higher Hurst parameter in comparison with a random baseline. The Hurst parameter can be used to accurately discriminate between the global dynamics of sociocultural systems (Gao, Fang, & Liu, 2017; Gao et al., 2012). Some signals show long-range memory (i.e., correlations at multiple time scale), while other signals only have short-range memory (i.e., correlation between neighbouring data points).

For trend reservoirs, Hurst exponent $H$ (i.e., an estimate of long-range dependencies), functions as an indicator variable (see equations in Appendix A). $H > 0.5$ indicates that a subreddit has long-range memory such that increases in a trend are followed by further increases, $H \approx 0.5$ indicates that a subreddit only displays short-range memory, likely due to a larger influx of diverse information, while $H < 0.5$ indicates anti-persistent and rigid behaviour (Gao, Cao, Tung, & Hu, 2007). Results in Nielbo, Vahlstrup, Gao, & Bechmann (2019) show that trend reservoirs have higher$H$ in comparison to the random sample - the higher the $H$ value the stronger the trend - and that $H$ has a stronger discriminatory power than the$\mathbb{N} * \mathbb{R}$ slope (i.e., it is better at detecting trend reservoirs).

Below, the classification of a subreddit as trending or not trending based on its long-range memory is exemplified with the subreddit *r/MachineLearning* (Figure 1).

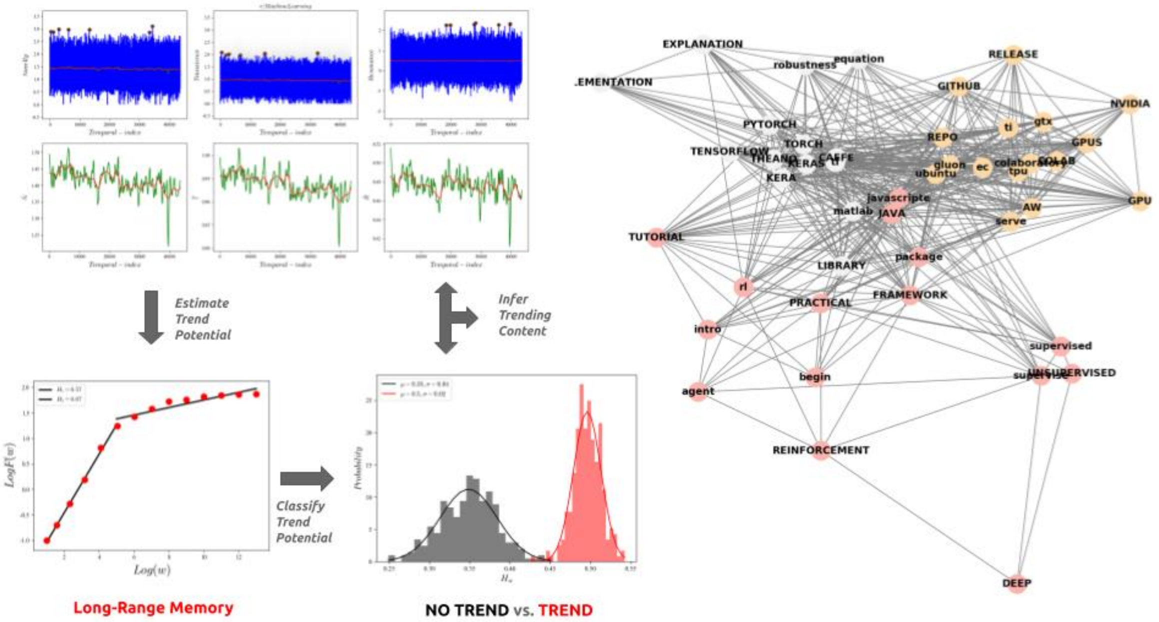

Figure 2: *Identification of trend potential as long-range memory on streaming data from r/MachineLearning on Reddit (find all figures in Appendix B)*

We start by computing information Novelty, Transience and Resonance for the subreddit (upper left corner in Figure 2 or Figure 2a in Appendix B). Then, we estimate the trend potential as the using the Hurst parameter (lower left corner in Figure 2 or Figure 2b in Appendix B).

Based on this estimation, we can then classify the subreddit as either trending or not trending on Reddit depending on the presence or absence of long-range memory (lower middle in Figure 2 and Figure 2c in Appendix B). With a Hurst parameter of $H = 0.57$, the *r/MachineLearning* subreddit can also be classified as trending as based on an estimate of its long-range dependencies. Finally, we infer the associated content on trending subreddits (graph on the right of Figure 2, see more in section 2.3).

Notice that these examples confirm that the Hurst parameter is better at discriminating between trending and not trending than the Novelty-Resonance slope. Comparing the distributions of $(\mathbb{N} * \mathbb{R})$ slopes and $H$ for random subreddits and the human-annotated trending subreddits, a larger overlap can be noticed for the distributions of $(\mathbb{N} * \mathbb{R})$ slopes than for the distributions of $H$ (Figure 3).

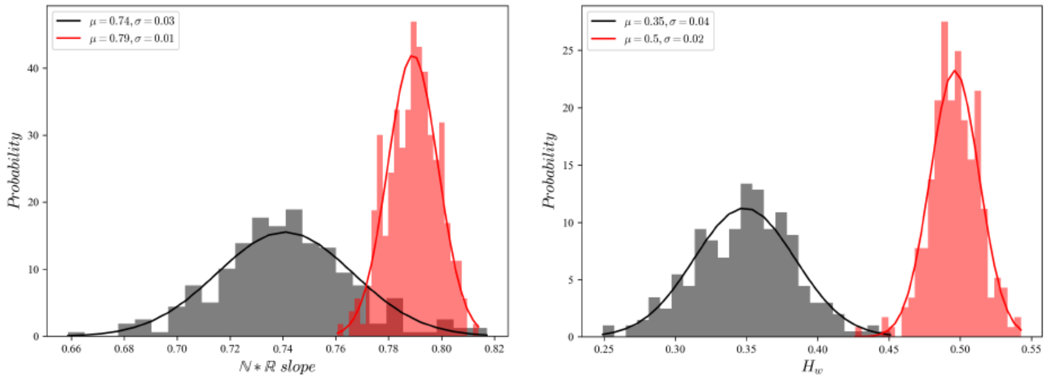

Figure 3: *Distributions of* $(\mathbb{N} * \mathbb{R})$ *slopes (left) and* $H$*(right) for random (gray) and trending (red).* $H$*has a stronger discrimination power than*$(\mathbb{N} * \mathbb{R})$ *for the two conditions.*

## 2.3 Content exploration: Extended search over the content graph

To facilitate content exploration, we train a simple neural embedding model (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) for each subreddit and query the model with the highest-ranking words in each word distribution $\varphi_k$ for topic *k*.

A simple neural embedding model is a shallow neural network that learns word association from the contexts of words. Basically, the model takes every word in the given data set, then takes the surrounding words (within a given window of words) of every word one-by-one and trains a neural network by feeding it all the word pairs found in the data set. The model can then predict the probability for each word to actually appear in the window around a given word. That means that instead of querying a data set for individual word occurrences, one can use the embedding model to query for the entire word context of any given query.

To build the concept graph of the word contexts, we compute the *m* associated terms which have the shortest arccosine distance to each query word, and then compute the *n* most associated terms using the same procedure for each of the *m* associated terms. We link the entire graph based on a distance threshold $\theta$ and detect communities of words using the Louvain method (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008).

The Louvain method is a simple, efficient and easy-to-implement method for detecting communities in large networks. The algorithm effectively detects groups of nodes within the network that are more densely connected to one another than to other nodes, and can thus help uncover topics within information networks (Fortunato & Castellano, 2007). In the context of our investigations, this method of content exploration of the subreddits may help us uncover the contexts in which certain technologies are discussed on Reddit.

# 3 Results

In the following, the results of our trend estimation model applied to discussions on Reddit (subreddits) are outlined. In this case, the model is used to discover and analyse the most trending subreddits related to artificial intelligence (AI), but the model can be applied to any given query represented as a seed list (see 2.1.1.). All figures and graphs can be found in Appendix B.

## 3.1 Identification of trend reservoirs in AI-related subreddits

Using the sampling criterion from section 2.1.1 we can sample the top 10 most relevant subreddits in two categories. The *Leaders* category contain subreddits with a substantial number of posts (more than 2560), while the *Prospects* category contain subreddits that can be small (only more than 256 posts) but rank the highest on the content matching.

Leaders are relevant because of the many posts of potentially trending content, while Prospects can be used to discover singular new trends. We can then use the trend reservoir estimation technique to classify into maximally trending and just trending or not trending on the assumption that not all subreddits that are in the trending sample are equally trending.

| Leaders in Reddit | | Prospects in Reddit | |
|---|---|---|---|
| *Subreddit* | *No. posts* | *Subreddit* | *No. posts* |
| r/technology | 1763527 | r/spectre_ai | 280 |
| r/futurology | 190990 | r/effectai | 592 |
| r/nvidia | 93118 | r/artificial | 20607 |
| r/stocks | 75315 | r/controlproblem | 1582 |
| r/MachineLearning | 66858 | r/aihub | 454 |
| r/mturk | 23638 | r/gameai | 783 |
| r/artificial | 20607 | r/wingsdao | 513 |
| r/learnmachinelearning | 12088 | r/bottos | 298 |
| r/deepdream | 10465 | r/dbrain | 499 |
| r/automate | 8965 | r/numerai | 265 |

Table 1: *Trend estimation of samples of AI related subreddits on Reddit.*

Table 1 displays the classification results, where both columns are the top 10 matches on the sampling criteria (i.e., the highest overlap between the description and a seed list of AI-related terms), but Leaders represents the largest subreddits that match the criteria, while Prospects is the best match to sample criteria.

The red subreddits are classified as maximally trending while blue are not. Leaders only qualify as maximally trending if they have a Novelty-Resonance slope steeper than 0.8 and display long-term memory with a Hurst parameter larger than 0.5. Prospects also qualify if they have either a Novelty-Resonance slope steeper than 0.8 or display long-term memory with a Hurst parameter larger than 0.5. This classification is extremely useful for decision support in terms of which subreddits to follow for a continuum of information on trends in e.g. AI.

Figure 4 exemplifies the procedure for classifying a subreddit as either maximally trending or just trending or not trending. In this case, the classification of subreddits is based on the Novelty-Resonance slope, but the classification procedure is the same for long-range memory.

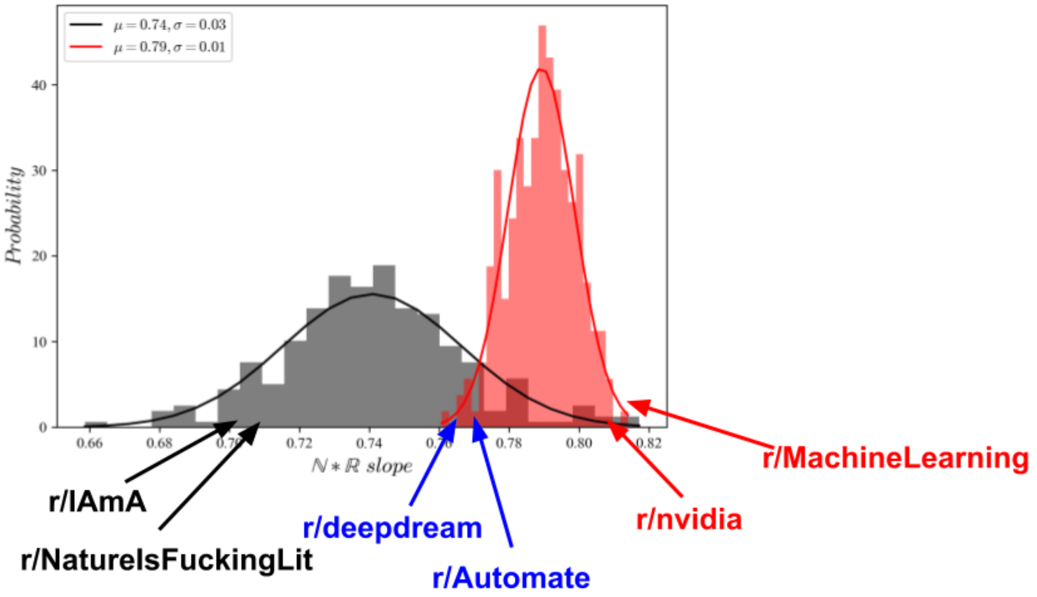

Figure 4: *Classification of subreddits based on the Novelty-Resonance slope. Black are random subreddits (gray distribution), blue are less trending subreddits in the trending sample (light red distribution), and red are maximally trending subreddits.*

Of the AI sample of subreddits in Table 1, some subreddits show trending potential (red) while others do not (blue). The comparison baseline is a random sample of subreddits (black). With this classifier, a recommender engine can easily identify trending subreddits within any given subject.

## 3.2 Topic content and Visualization in *r/MachineLearning*

Once the subreddits are classified as trending or not, we can deep dive into the maximally trending subreddits by querying the neural embedding model with the topic model and depicting the results as a concept graph (see 2.3).

Figure 5 (next page) shows the concept graph we call TOOL & MOTIVATION from the maximally trending subreddit r/MachineLearning. The overall structure of the graph is split in two conceptual clusters “Tools of the Trade” (clustering, classification, decoder, LSTM, RNN) and “Competitions” (e.g., Netflix, prize, Kaggle, Amazon, contest) with the grey nodes representing tasks that bind together the two conceptual clusters.

This concept graph spans a space that on the one hand shows necessary knowledge (Tools of the Trade) to work within the leading paradigm of machine learning: Deep learning and deep neural networks; and on the other hand, a central motivational component (Competitions): Data science competitions and hackathons. The nodes that mediate between “Tools of the Trade” and “Competitions” (e.g., charact(er), digit(te) and (object) recognition (yolo)) represent the tasks that often bind together tools and competitions.

“Tools of the Trade” also display interesting sub-clusters. In the upper part, the graph displays classes of problems (classification, clustering), towards the centre elements related to data (text, image, noisy, binary, variation), and to the right classes of deep neural networks (autoencoder, CNN, RNN, LSTM).

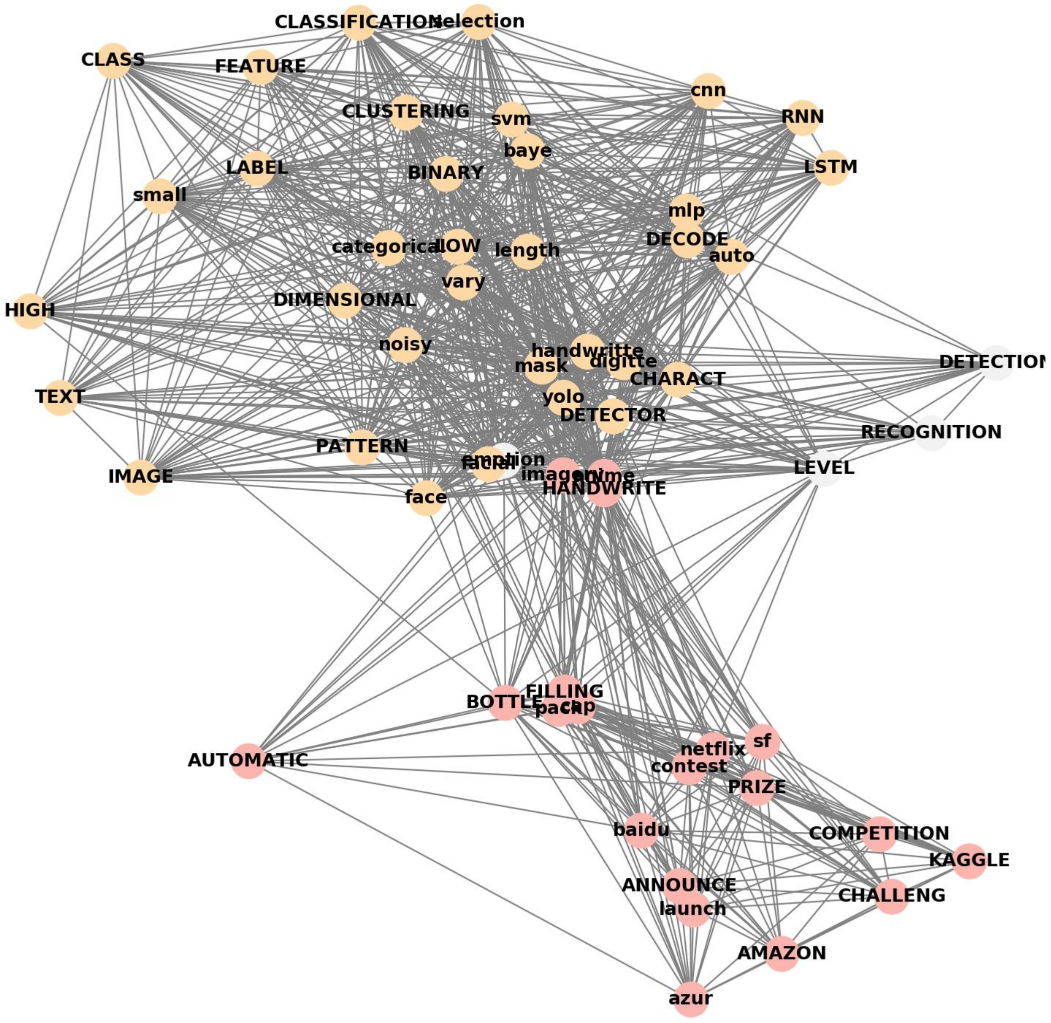

Figure 5: *TOOL & MOTIVATION concept graph of r/MachineLearning*

We call the concept graph in Figure 6 TOOL-DIVERSIFICATION (next page). Data science and machine learning is a complicated field and many classes of tools are necessary to develop state-of-the-art deep learning models, and the concept graph shows three clusters of interest:

- The upper right corner shows important tools related to “Hardware and Cloud” technologies (NVIDIA, GPU, TPU, AWS, server, Ubuntu, Gluon), which are all characteristics of GPU accelerated high performance computing;
- The cluster in the centre left of the graph is dominated by the most important deep learning “Software Libraries” in Python (Tensorflow, PyTorch, Keras, Theano) and related languages (JavaScript, Java, Caffe MATLAB);
- In the lower part, the graph displays “Classes of Problems” (supervised, unsupervised, reinforcement learning).

Two further observations can be made. Firstly, “Tutorial” is highly interconnected to all clusters, supporting the fact that tutorials have become one of the primary sources of assimilating the diverse tools in the machine learning community. Secondly, software libraries, packages, and frameworks take a central role in the graph. They all signify bundles of pre-existing code that minimize the amount of programming and hardware understanding required by machine learning enthusiasts.

These observations indicate that the subreddit does not consist of solely professional machine learning developers, but rather constitutes a community of machine learning enthusiasts with a do-it-yourself approach to machine learning.

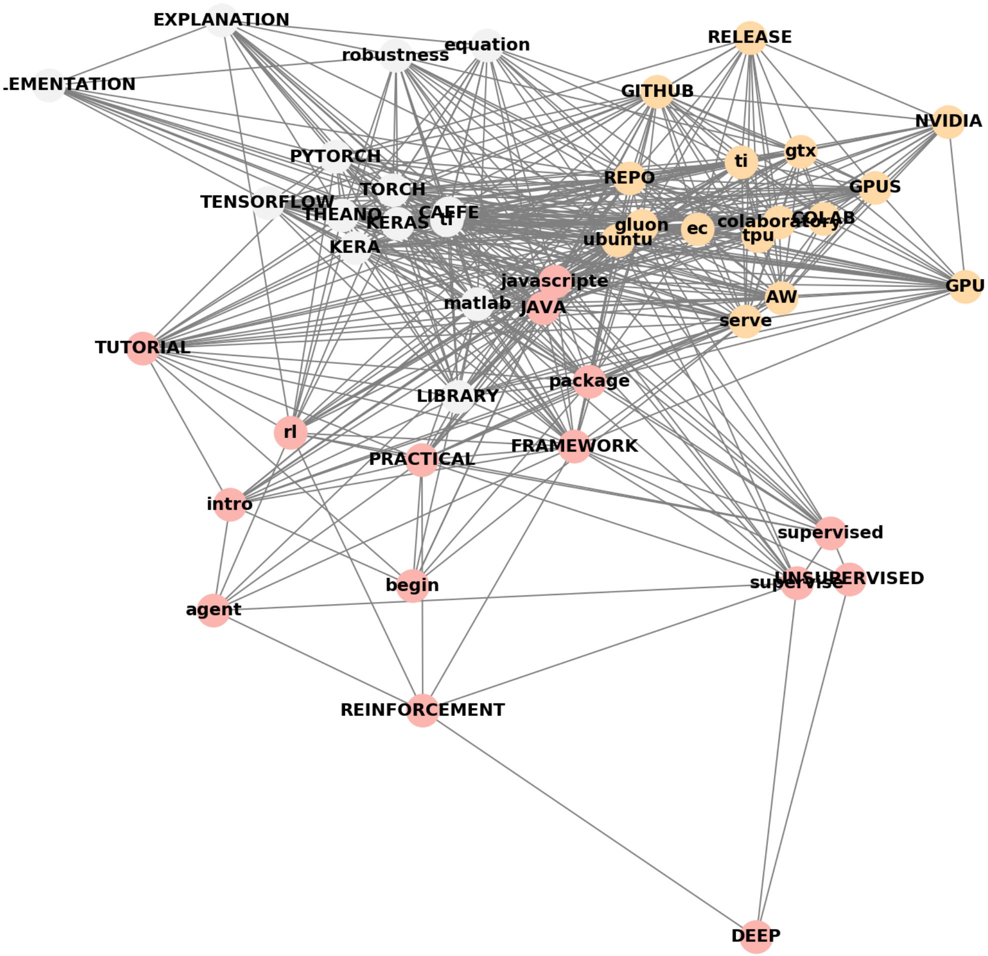

Figure 6: *TOOL-DIVERSIFICATION concept graph of r/MachineLearning*

The concept graph in Figure 7 (next page) represent discussions related to GPU accelerated computing in the trending subreddit *r/nvidia*. GPU accelerated high performance computing has taken central stage in the development and applications of artificial intelligence, and Nvidia is the main manufacturer of GPU accelerators.

Two hardware classes are exemplified by the graph: Relatively cheap commodity GeForce cards developed for computer graphics in grey (GTX, TI, TIs, RTX) and professional Tesla cards for scientific applications in red (Pascal, Volta, Maxwell architectures). The cluster in yellow is shared between both classes of hardware and reflect performance measures of accelerators that are used to identify and build the relevant computing infrastructure.

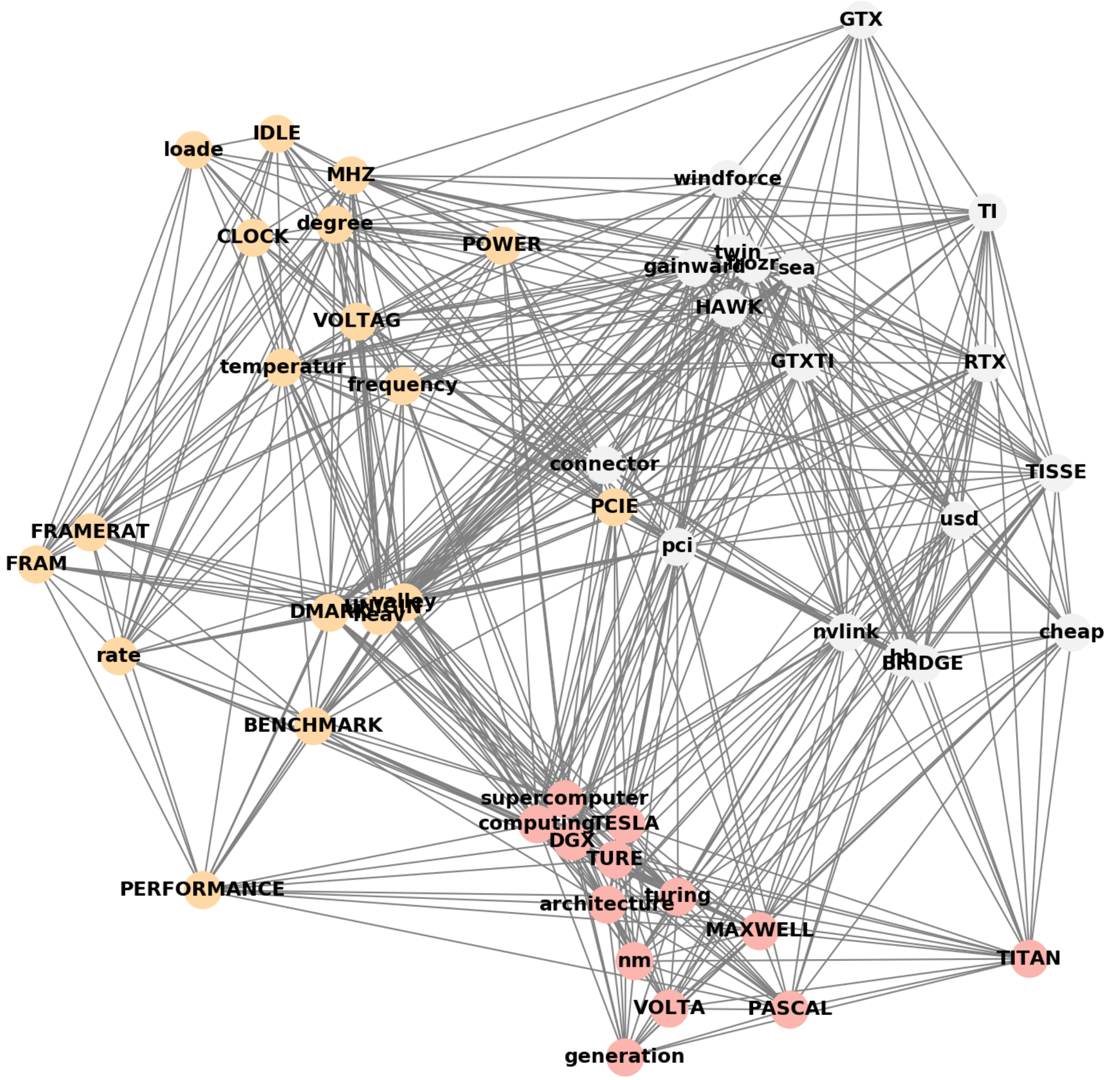

Figure 7: *GPU-ACCELERATED COMPUTING concept graph of r/nvidia*

In Appendix B, you can find concepts graphs from other trending AI-related subreddits *r/futurology* and *r/technology*. These and the graphs above are just examples of content that can be extracted after identifying the most trending subreddits within the topic of investigation.

# 4 Conclusion

This report presented a new approach to trend estimation on social media that identifies trend reservoirs according to their relationship between Novelty and Resonance, and their degree of persistence. Results show that trend reservoirs have steeper $\mathbb{N} * \mathbb{R}$ slope and higher $H$ in comparison to a random baseline. Importantly, these 'signatures' capture different properties of trend reservoirs, information stickiness and multi-scale correlations respectively, and they both have discriminatory power, i.e. they can actually detect trend reservoirs.

The model for trend estimation was applied to discover trending discussions related to internet development on the social media platform Reddit. In this report, the model was used for a deep dive into discussions related to artificial intelligence, discovering the most trending subreddits from a sample of subreddits with the highest overlap between the description and a seed list of AI-related terms.

A recommender engine can be trained with this classifier to identify trending subreddits within any given subject. Such classifications can be extremely useful for decision support in terms of which subreddits to follow for a continuum of information on trends in e.g. AI.

To exemplify how content on trending subreddits can be explored after the classification, a neural embedding model was trained to query the highest-ranking words and their associated words. This approach provides concepts graphs of the discussions in the subreddit that gives insights into e.g. the contexts in which certain technologies are discussed.

These are solely the intermediary results of the social media analysis, detailing the development, testing and confirmation of the new model for trend estimation developed for the project, and exemplifying the application of the model on Reddit data.

The model will be further applied to investigate internet-related discussions on social media in *D1.8 Final social media analysis report & visualisations* to be delivered in 2021, which will also include an analysis of internet-related discussions on Twitter.

The model will also be used to inform the official selection of NGI topics and the topic guides as part of *D1.9 NGI Topic guides & evaluation criteria report I* to be delivered in March, 2020.

# 5 References

Barron, A. T. J., Huang, J., Spang, R. L., & DeDeo, S. (2018). Individuals, institutions, and innovation in the debates of the French Revolution. *Proceedings of the National Academy of Sciences*, *115*(18), 4607–4612. https://doi.org/10.1073/pnas.1717729115

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Chinnov, A., Kerschke, P., Meske, C., Stieglitz, S., & Trautmann, H. (2015). An Overview of Topic Discovery in Twitter Communication through Social Media Analytics. *AMCIS*, 10.

Fortunato, S., & Castellano, C. (2007). Community Structure in Graphs. *ArXiv:0712.2716 [Cond-Mat, Physics:Physics]*. Retrieved from http://arxiv.org/abs/0712.2716

Gao, J., Cao, Y., Tung, W., & Hu, J. (2007). *Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond* (1 edition). Hoboken, N.J: Wiley-Interscience.

Gao, J., Fang, P., & Liu, F. (2017). Empirical scaling law connecting persistence and severity of global terrorism. *Physica A: Statistical Mechanics and Its Applications*, *482*(C), 74–86.

Gao, J., Hu, J., Mao, X., & Perc, M. (2012). Culturomics meets random fractal theory: Insights into long-range correlations of social and natural phenomena over the past two centuries. *Journal of The Royal Society Interface*, *9*(73), 1956–1964. https://doi.org/10.1098/rsif.2011.0846

Gao, J., Jockers, M. L., Laudun, J., & Tangherlini, T. R. (2016). A multiscale theory for the dynamical evolution of sentiment in novels. *2016 International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC)*, 1–4. https://doi.org/10.1109/besc.2016.7804470

Gray, K. (2007). Comparison of Trend Detection Methods. *Graduate Student Theses, Dissertations, & Professional Papers*. Retrieved from https://scholarworks.umt.edu/etd/228

Gyódi, K., Nawaro, Ł., Paliński, M., & Wilamowski, M. (2019). *D1.2: Visualisations of key emerging technologies and social issues*. The European Commission.

Horne, B. D., Adali, S., & Sikdar, S. (2017). Identifying the social signals that drive online discussions: A case study of Reddit communities. *ArXiv:1705.02673 [Cs]*. Retrieved from http://arxiv.org/abs/1705.02673

Hu, Q., Liu, B., Thomsen, M. R., Gao, J., & Nielbo, K. L. (2019). Dynamic evolution of sentiments in Never Let Me Go: Insights from multifractal theory and its implications for literary analysis. *HAL Preprint Hal-02143896*.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, 591. https://doi.org/10.1145/1772690.1772751

Madani, A., Boussaid, O., & Zegour, D. E. (2014). What's Happening: A Survey of Tweets Event Detection. *ICC 2014*.

Mathioudakis, M., & Koudas, N. (2010). TwitterMonitor: Trend detection over the twitter stream. *Proceedings of the 2010 International Conference on Management of Data - SIGMOD '10*, 1155. https://doi.org/10.1145/1807167.1807306

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. s, & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Murdock, J., Allen, C., & DeDeo, S. (2017). Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks. *Cognition*, *159*, 117–126. https://doi.org/10.1016/j.cognition.2016.11.012

Nielbo, Kristoffer L., Perner, M. L., Larsen, C., Nielsen, J., & Laursen, D. (2019). Automated compositional change detection in Saxo Grammaticus' Gesta Danorum. *CEUR*

*Workshop Proceedings*, *2364*, 321–332. Retrieved from https://portal.findresearcher.sdu.dk/en/publications/automated-compositional-change-detection-in-saxo-grammaticus-gest

Nielbo, Kristoffer L., Vahlstrup, P. B., Gao, J., & Bechmann, A. (2019). *Sociocultural trend signatures in minimal persistence and past novelty*. Manuscript submitted for publication.

Nielbo, Kristoffer Laigaard, Baunvig, I. K. F., Liu, B., & Gao, J. (2019). A curious case of entropic decay: Persistent complexity in textual cultural heritage. *Digital Scholarship in the Humanities*, *34*(3). https://doi.org/10.1093/llc/fqy054

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Tenen, D. Y. (2018). Toward a Computational Archaeology of Fictional Space. *New Literary History*, *49*(1), 119–147. https://doi.org/10.1353/nlh.2018.0005

Wevers, M., Gao, J., & Nielbo, K. L. (2019). Tracking the Consumption Junction: Temporal Dependencies between Articles and Advertisements in Dutch Newspapers. *ArXiv*, *abs/1903.11461*.

# 6 Appendices

## 6.1 Appendix A: Equations

Appendix A details the equations involved in the new model for trend estimation.

### Novelty and Resonance

For estimates of Novelty and Resonance, a Latent Dirichlet Allocation (LDA) model was trained for each subreddit in order to create dense low-rank vector representations (Blei, Ng, & Jordan, 2003). A grid search was carried out for each model in order to determine the parameter $K$ (number of topics) from 10 to 250 in steps of 10 and the loglikelihood of each model was used as an evaluation metric. Novelty ($\mathbb{N}$), Transience ($\mathbb{T}$) and Resonance ($\mathbb{R}$) were estimated for a window ($w$) of three days and based on the following equations from Barron, Huang, Spang, & DeDeo (2018):

$$\mathbb{N}_w(j) = \frac{1}{w}\sum_{d=1}^{w} D_{KL}\left(s^{(j)} \mid s^{(j-d)}\right)$$

$$\mathbb{T}_w(j) = \frac{1}{w}\sum_{d=1}^{w} D_{KL}\left(s^{(j)} \mid s^{(j+d)}\right)$$

$$\mathbb{R}_w(j) = \mathbb{N}_w(j) - \mathbb{T}_w(j)$$

Where s is a $K$-dimensional document distribution in the LDA model and $D_{KL}$ is the Kullback-Leibler divergence:

$$D_{KL}(s^{(j)} \mid s^{(k)}) = \sum_{i=1}^{K} s_i^{(j)} \times log_2 \frac{s_i^{(j)}}{s_i^{(k)}}$$

Because LDA can give less than optimal results for short documents, the performance of each model was compared to a model trained on the same data using Non-negative Matrix Factorization and cosine distance (Moyer, Dye, Carson, Carson, & Goldbaum, 2015). Signal properties were robust across models and LDA chosen for continuity with previous studies.

### Nonlinear Adaptive Filtering

Nonlinear adaptive filtering is used because of the inherent noisiness of trend signals (Gray, 2007). First, the signal is partitioned into segments (or windows) of length $w = 2n + 1$points, where neighboring segments overlap by $n + 1$. The time scale is $n + 1$ points, which ensures symmetry. Then, for each segment, a polynomial of order $D$ is fitted. Note that $D = 0$ means a piece-wise constant, and $D = 1$ a linear fit. The fitted polynomial for $ith$ and $(i + 1)th$ is denoted as $y^{(i)}(l_1), y^{(i+1)}(l_2)$, where $l_2, l_2 = 1,2,\ldots,2n + 1$. Note the length of the last segment may be shorter than w. We use the following weights for the overlap of two segments.

$$y^{(c)}(l_1) = w_1 y^{(i)}(l+n) + w_2 y^{(i)}(l), l = 1,2,\dots,n+1$$

Where $w_1 = (1-\frac{l-1}{n}), w_2 = 1 - w_1$ can be written as $(1-\frac{d_1}{n}), j = 1,2,$ where $d_j$ denotes the distance between the point of overlapping segments and the center of $y^{(i)}, y^{(i+1)}$. The weights decrease linearly with the distance between point and center of the segment. This ensures that the filter is continuous everywhere, which ensures that non-boundary points are smooth.

**Adaptive Fractal Analysis**

Assuming that stochastic process $X = X_t : t = 0,1,2,\dots$, with stable covariance, mean $\mu$ and $\sigma^2$, the process' autocorrelation function for $r(k), k \geq 0$ is:

$$r(k) \ = \frac{E[X(t)x(t+k)]}{E[X(t)^2]} \sim k^{2H-2}, as \quad k \to \infty$$

Where $H$ is called the Hurst parameter (Mandelbrot, 1982). For $0.5 > H > 1$ the process is characterized by long-range temporal correlations such that increments are followed by increases and decreases by further decreases. For $H = 0.5$ the time series only has short-range correlations; and when $H > 0.5$ the time series is anti-persistent such that increments are followed by decreases and decreases by increments.

Detrended fluctuation analysis (DFA) is the most widely used method for estimating the Hurst parameter, but DFA may involve discontinuities at the boundaries of adjacent segments. Such discontinuities can be detrimental when the data contain trends (K. Hu, Ivanov, Chen, Carpena, & Eugene Stanley, 2001), non-stationarity (Kantelhardt et al., 2002), or nonlinear oscillatory component (Chen et al., 2005; J. Hu, Gao, & Wang, 2009). Adaptive fractal analysis (AFA) is a more robust alternative to DFA (Gao, Hu, & Tung, 2011; Tung, Gao, Hu, & Yang, 2011). AFA consists of the following steps: first, the original process is transformed to a random walk process through first-order integration:

$$u(n) = \sum_{k=1}^{n} (x(k) - \underline{x}), N = 1,2,3,\dots,N,$$

where $\underline{x}$ is the mean of $x(k)$. Second, we extract the global trend $(v(i), i = 1,2,3,\dots,N)$ through the nonlinear adaptive filtering. The residuals $(u(i) - v(i))$ reflect the fluctuations around a global trend. We obtain the Hurst parameter by estimating the slope of the linear fit between the residuals' standard deviation $F^{(2)}(w)$ and $w$ window size as follows:

$$F^{(2)}(w) = \left[\frac{1}{N}\sum_{i=1}^{N}(u(i) - v(i))^2\right]^{\frac{1}{2}} \sim w^H$$

## 6.1 Appendix B: Figures

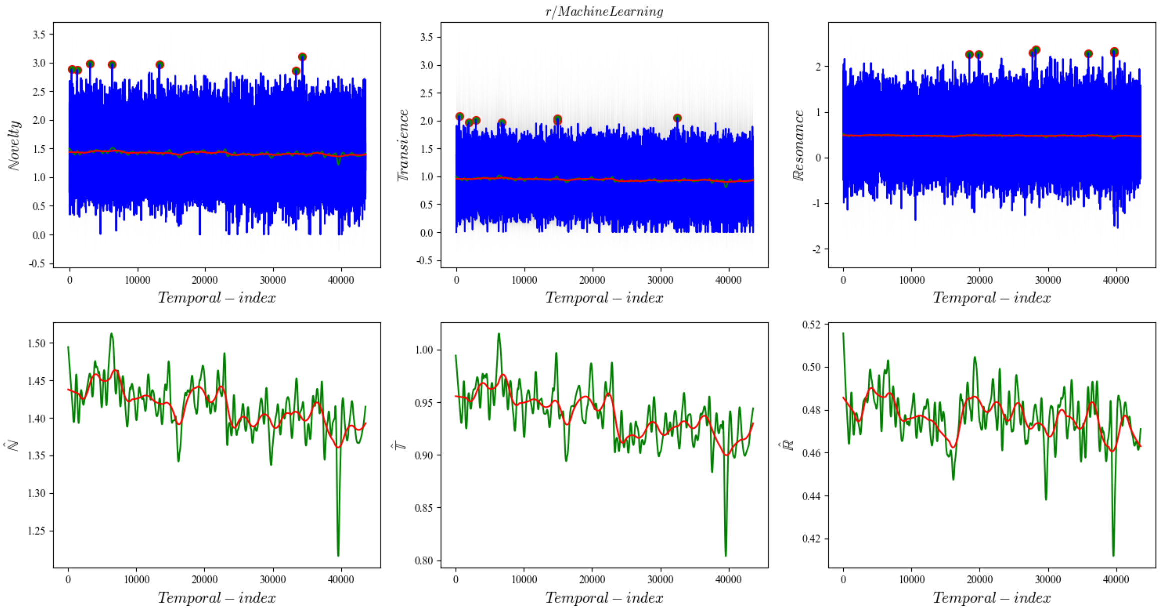

Figure 1a and 2a: *Novelty, Transience and Resonance for r/MachineLearning*

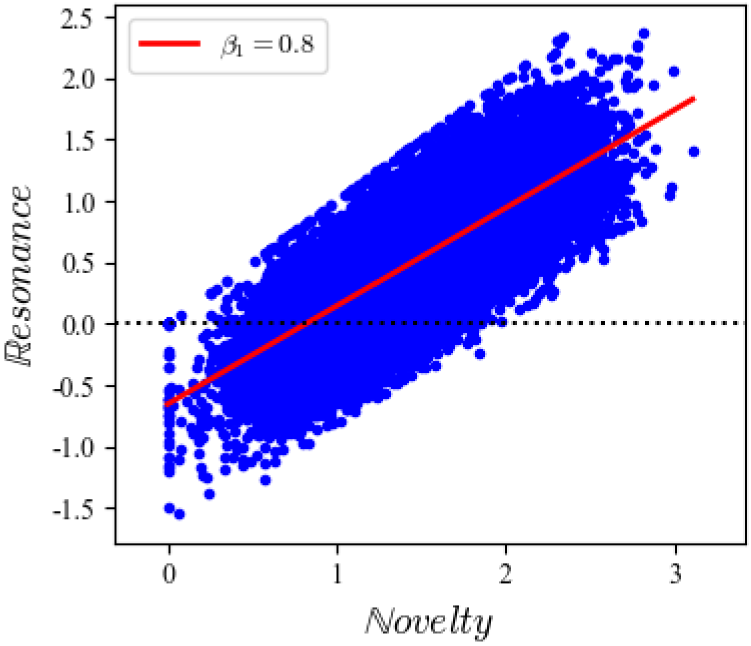

Figure 1b: *Novelty-Resonance Slope for r/MachineLearning*

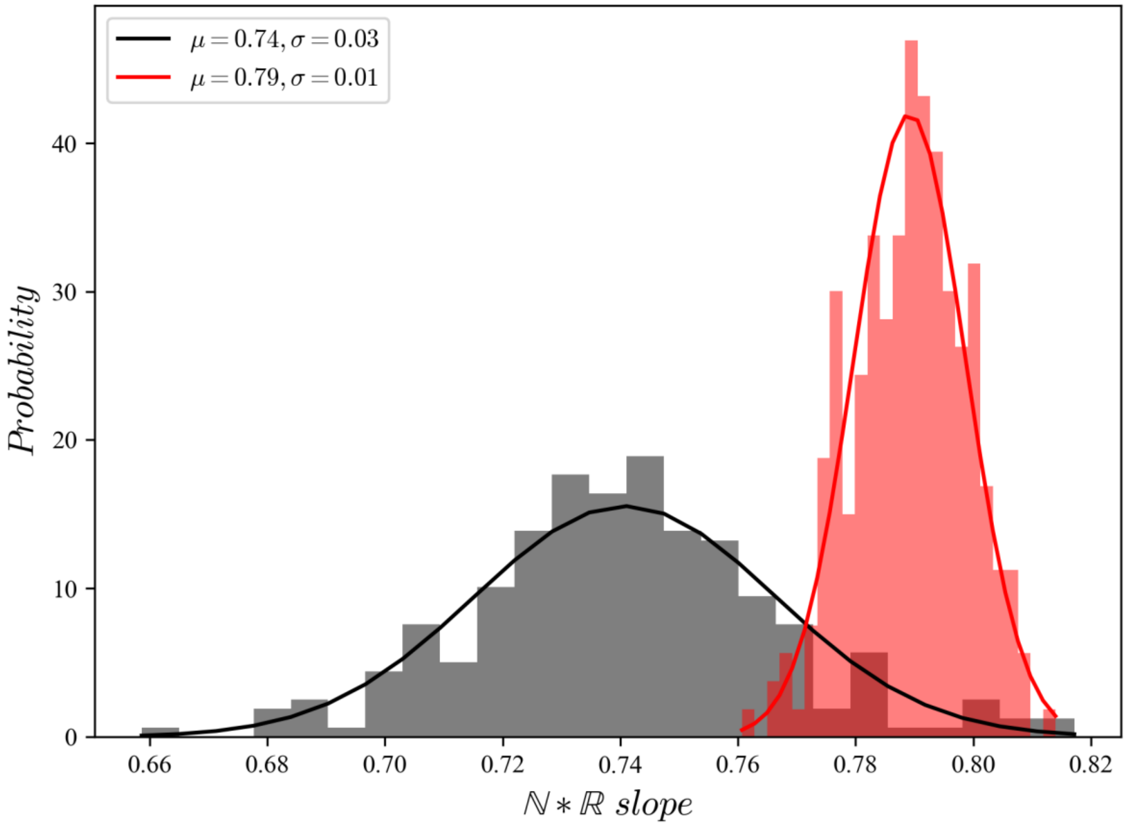

Figure 1c: *NO TREND vs. TREND classified on* $\mathbb{N} * \mathbb{R}$ *for r/MachineLearning*

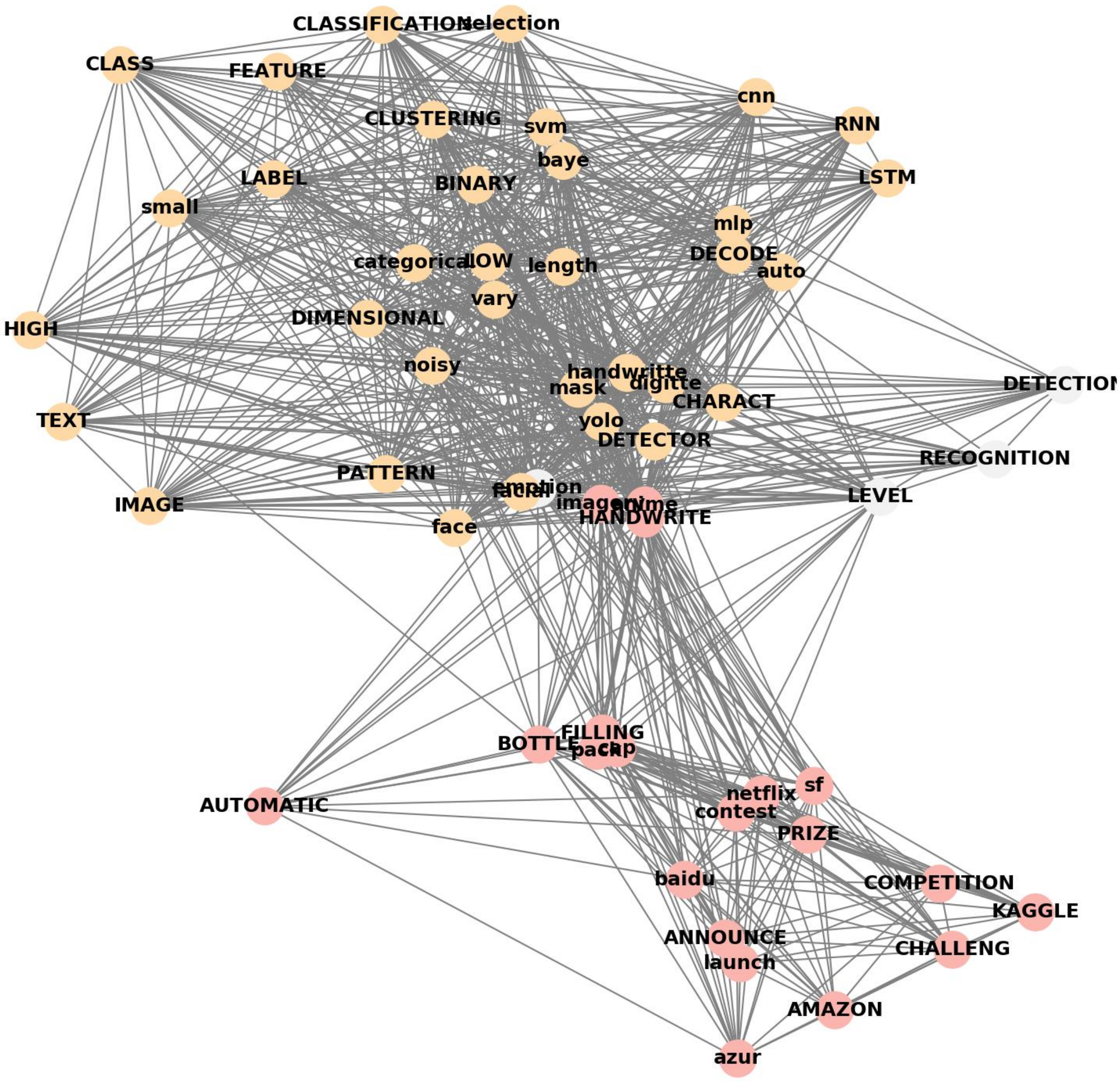

Figure 1d & 5: *Associated content on r/MachineLearning*

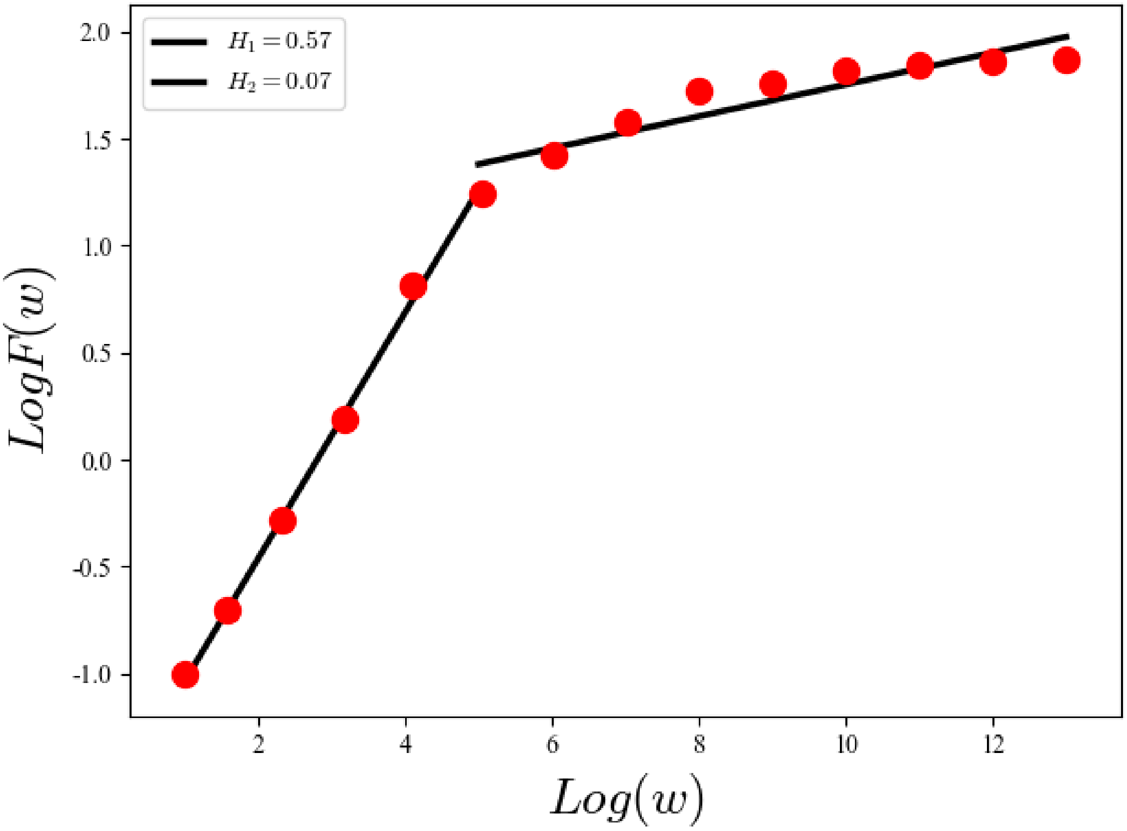

Figure 2b: Estimation of *Long-range memory for r/MachineLearning*

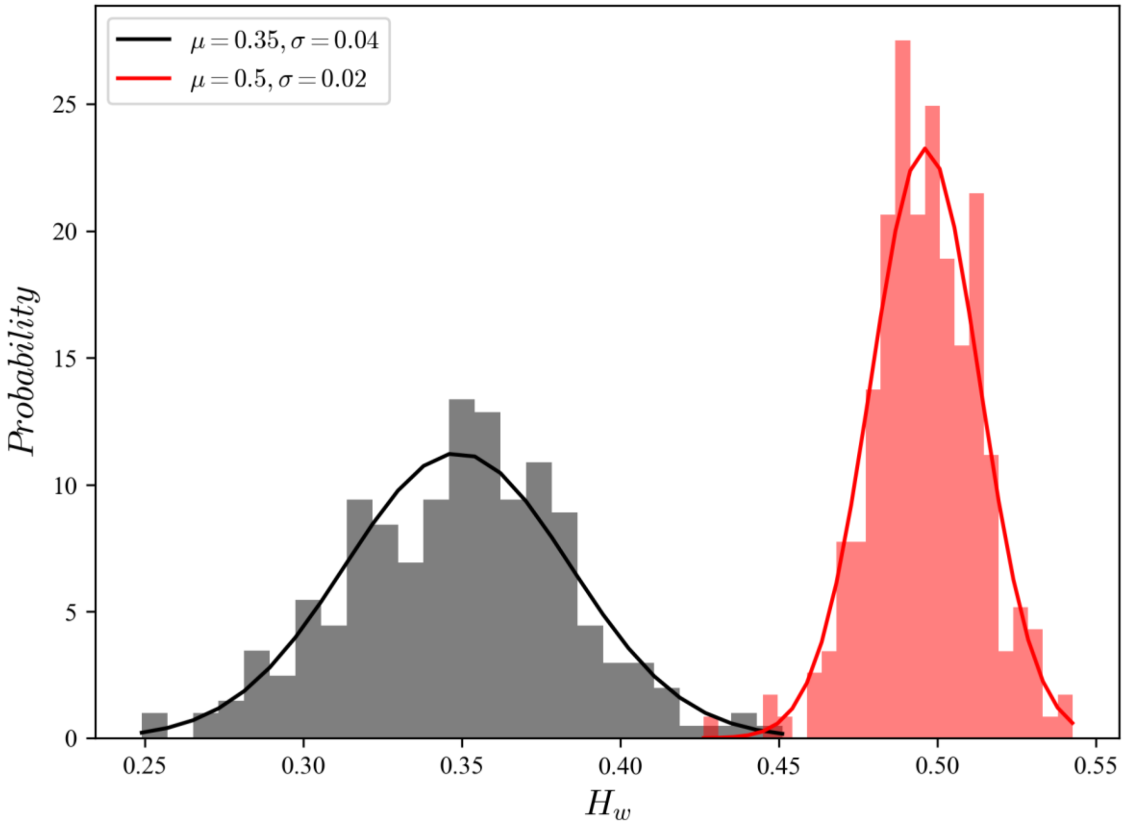

Figure 2c: *NO TREND vs. TREND* classified on long-range memory (*H*) *for r/MachineLearning*

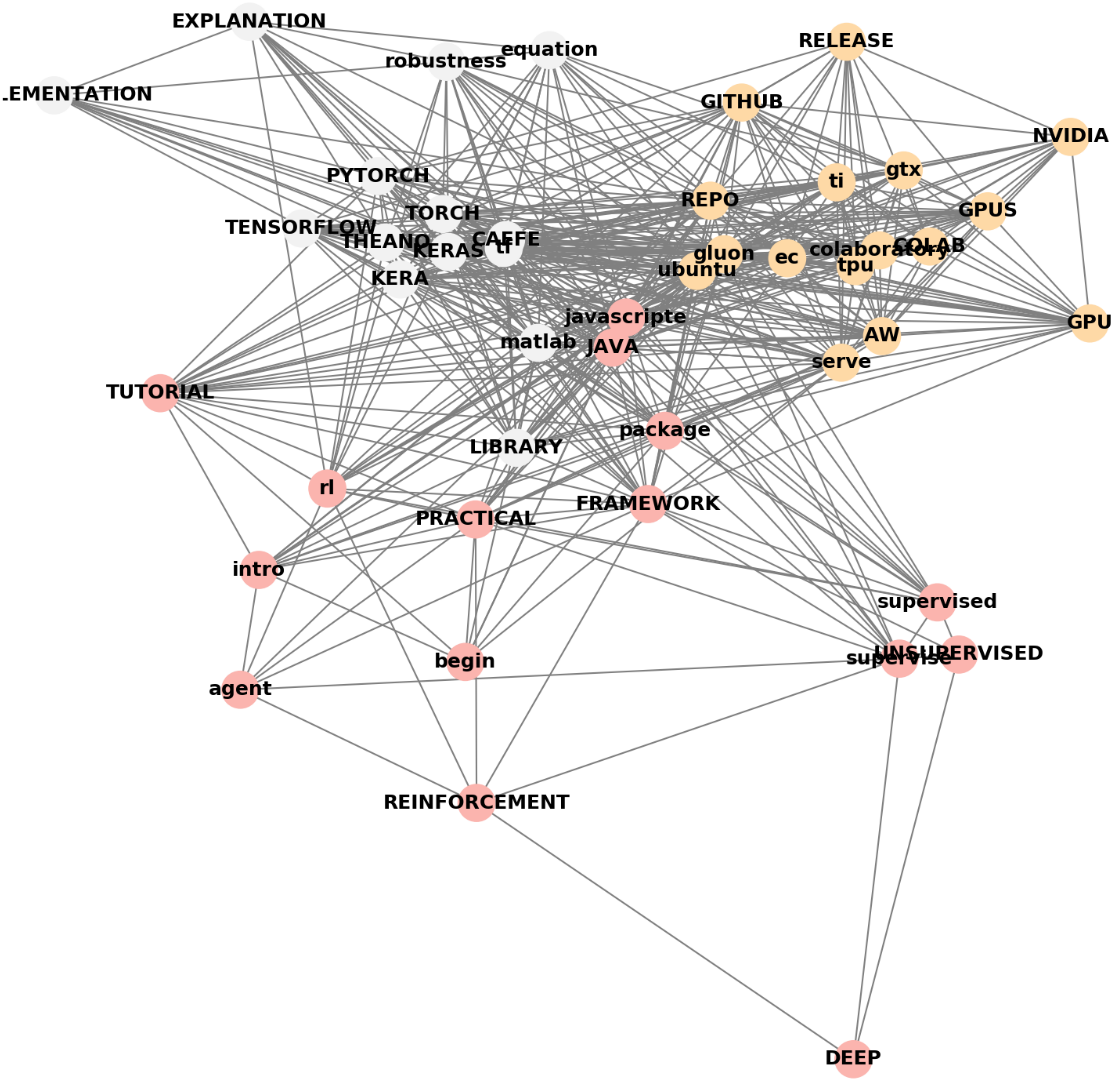

Figure 2d & 6: *Associated content on r/MachineLearning*

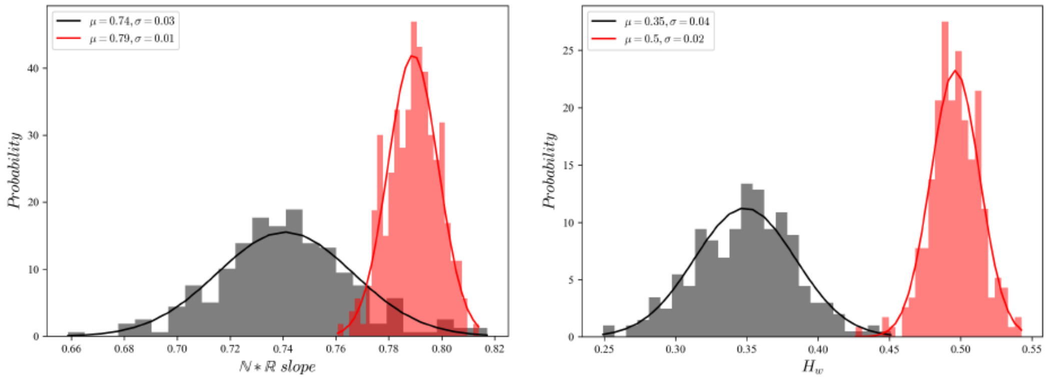

Figure 3: *Distributions of* $(\mathbb{N} * \mathbb{R})$ *slopes (left) and* $H$*(right) for random (gray) and trending (red).*

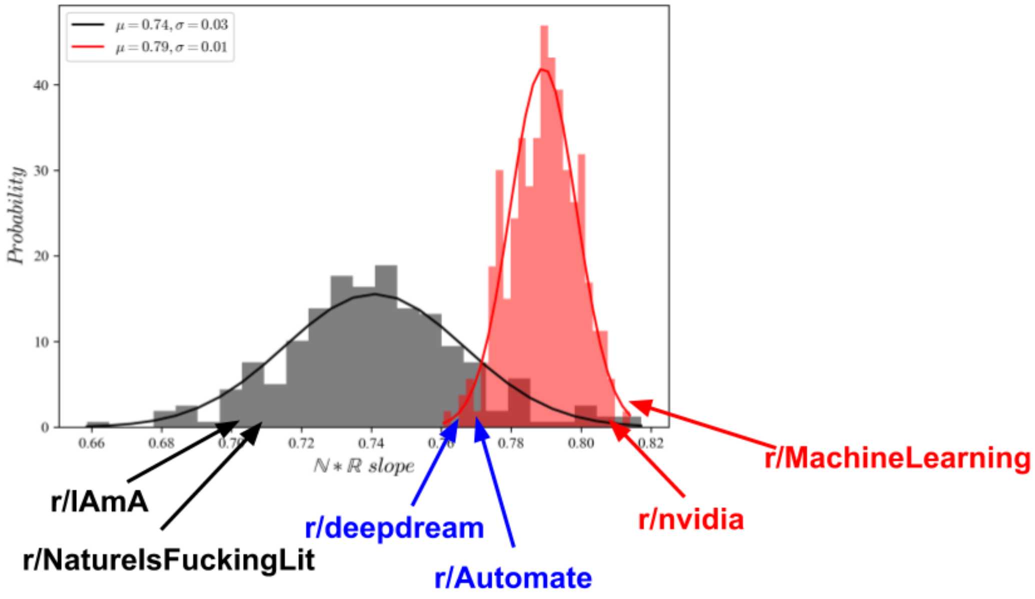

Figure 4: *Classification of subreddits based on the Novelty-Resonance slope*

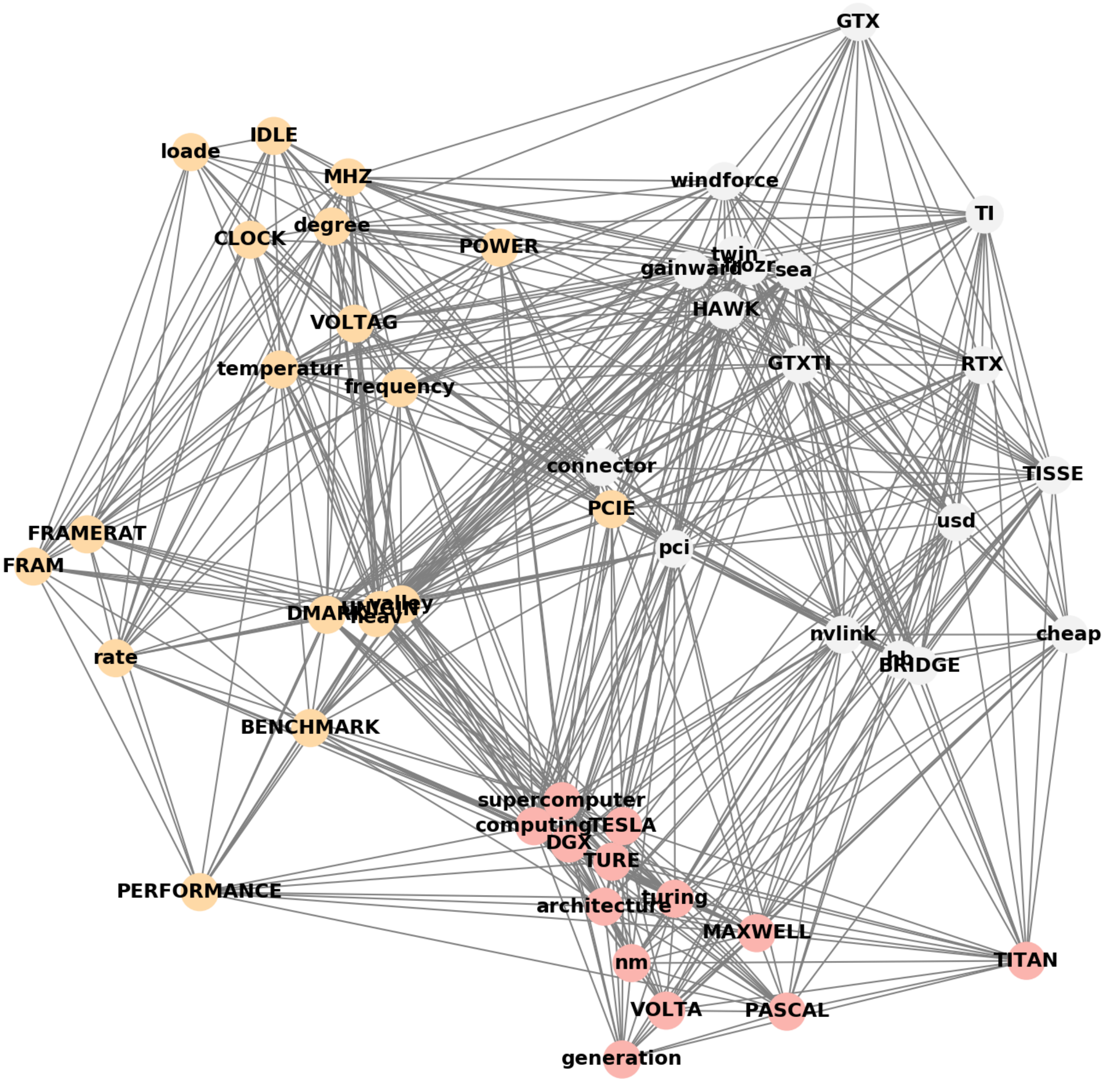

Figure 7: *GPU-ACCELERATED COMPUTING concept graph of r/nvidia*

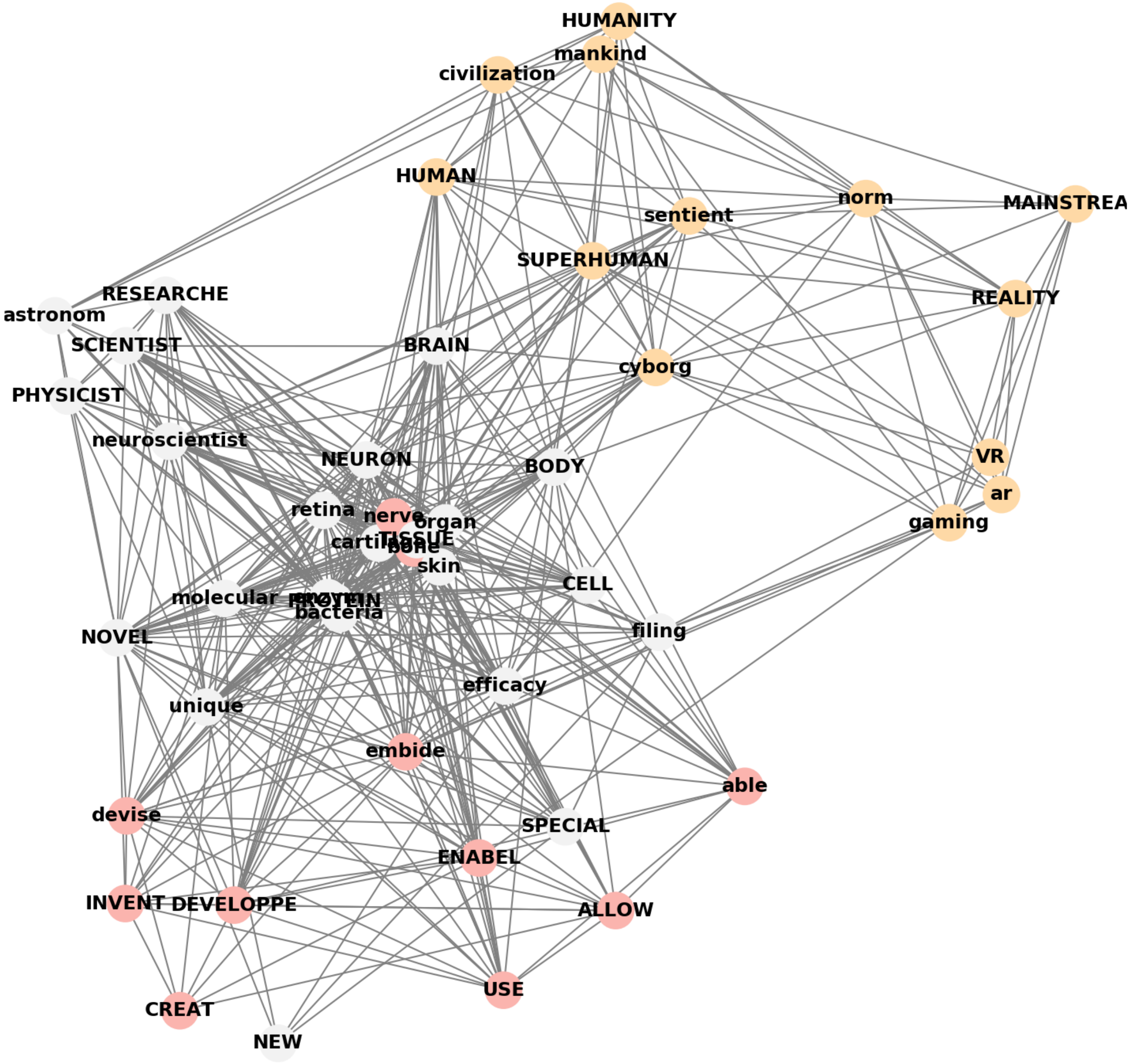

Figure 8: *Concept graph from r/futurology*

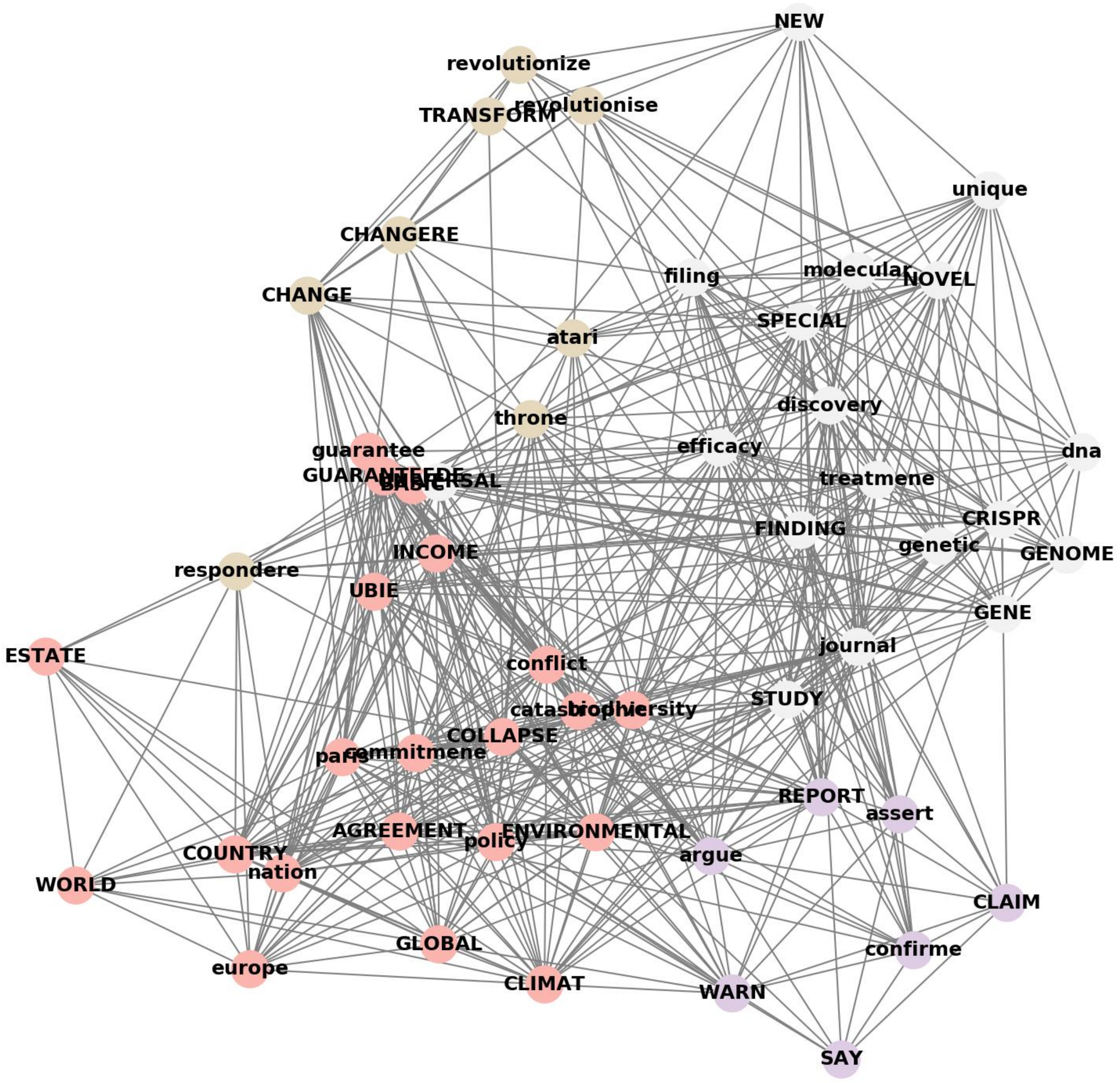

Figure 9: *Concept graph from r/futurology*

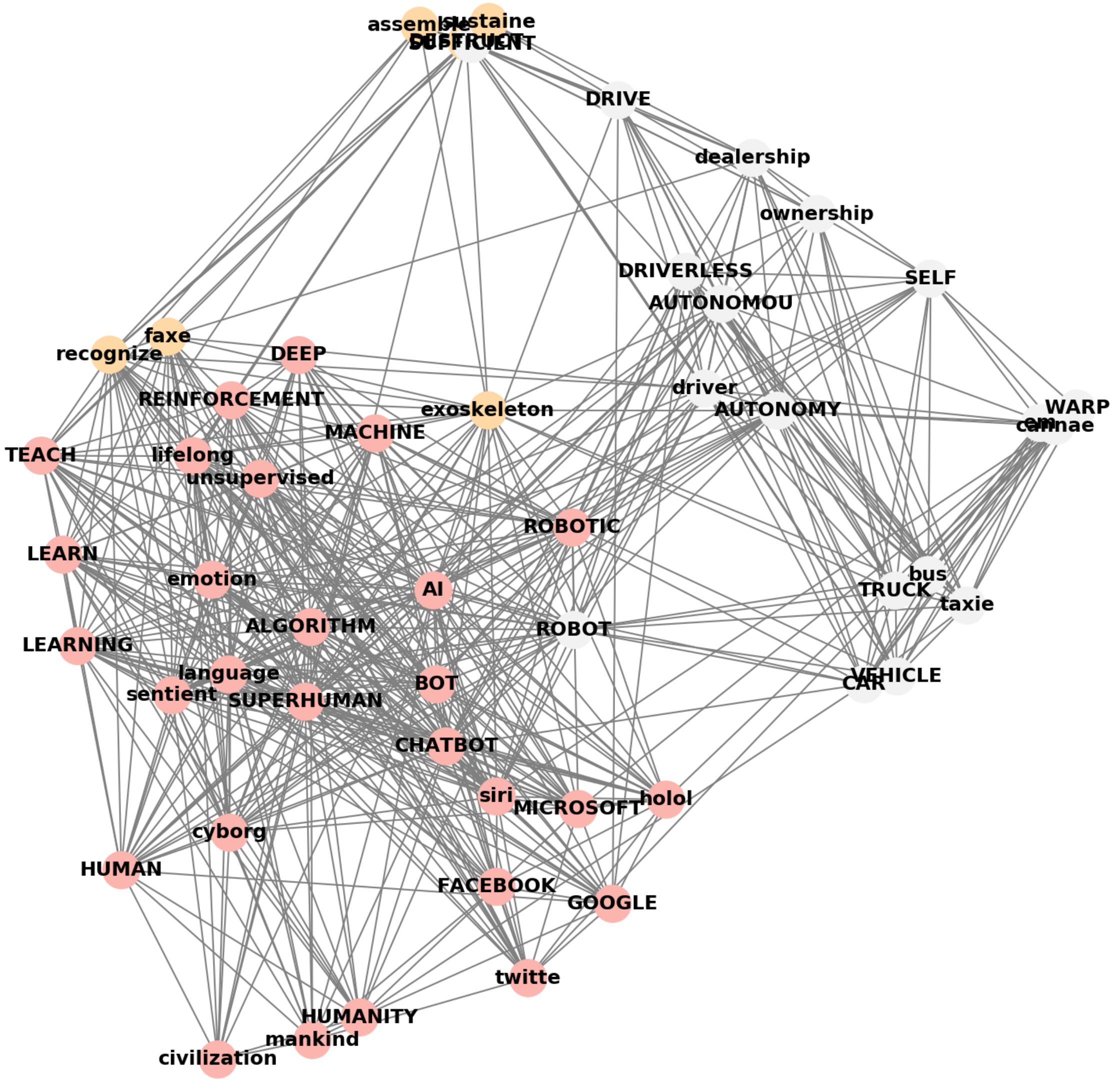

Figure 10: *Concept graph from r/futurology*

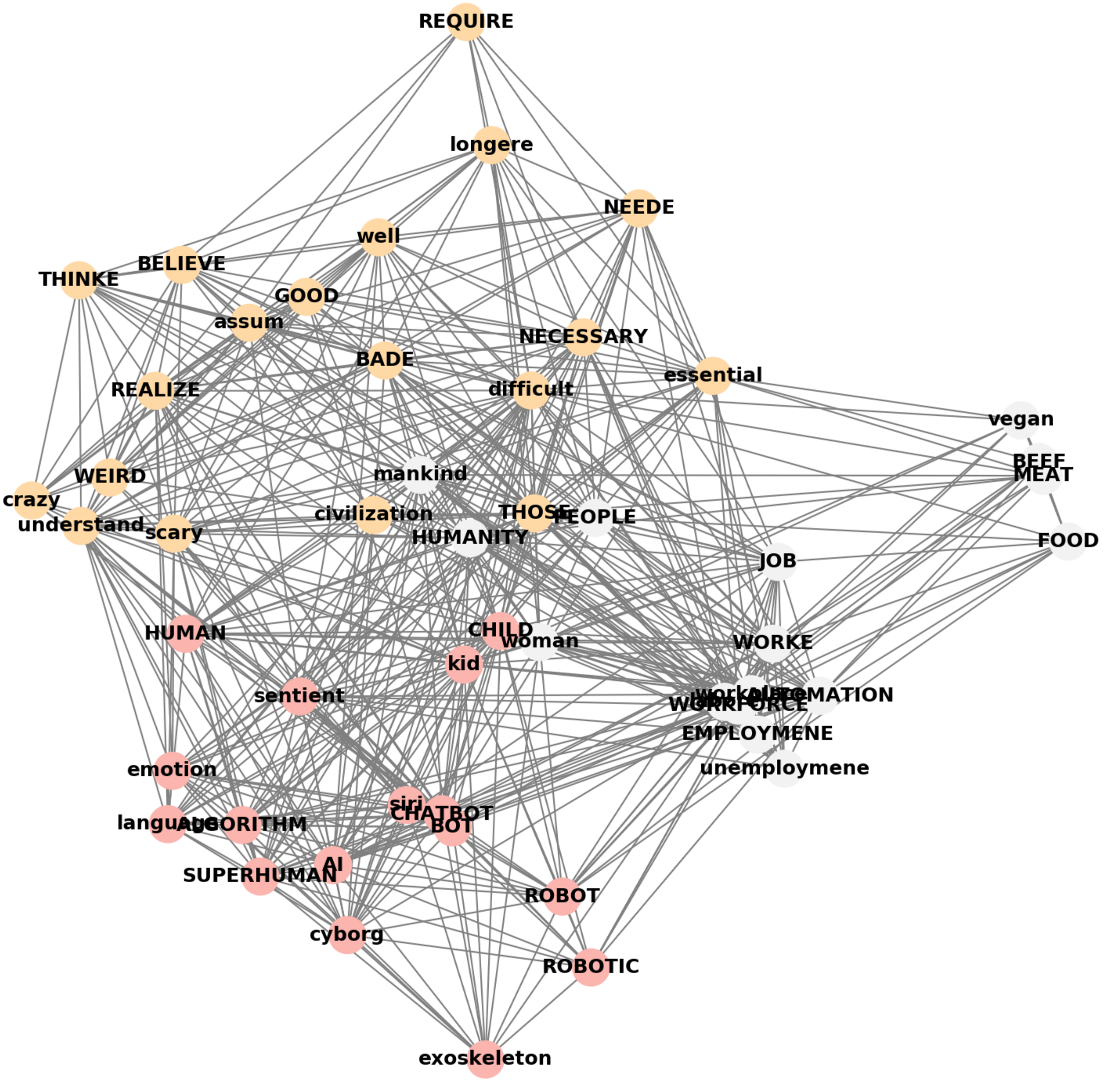

Figure 11: *Concept graph from r/futurology*

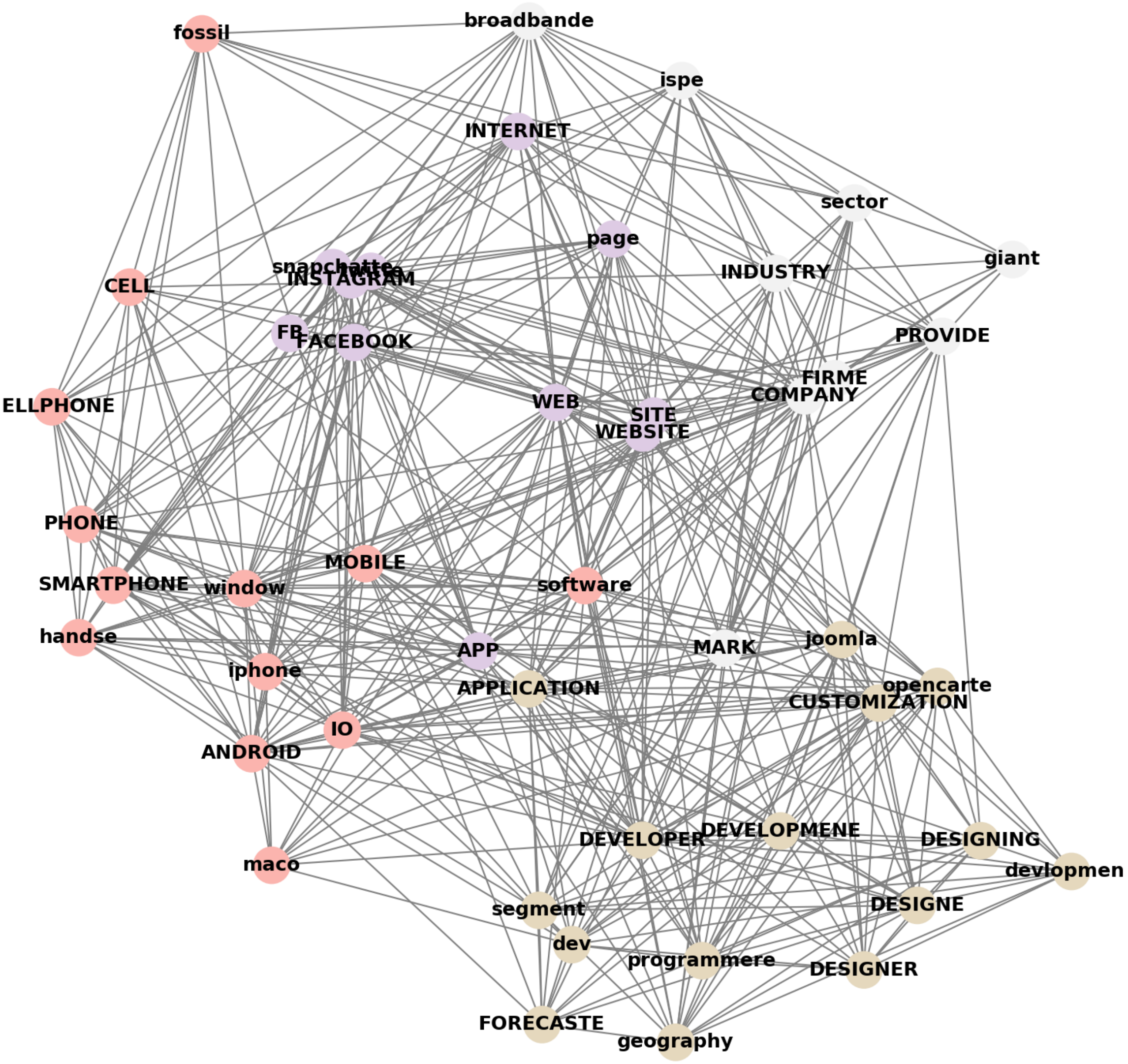

Figure 12: *Concept graph from r/technology*