ISyE 6740 - Spring 2021
Final Report

Team Member Names:	Koh Rui Nah
Project Title:	Finding the hottest topics in Tech

Please include (at least) the following sections.

Problem Statement

Technology (Tech) is a domain in a constant flux. New technology and terms have emerged rapidly over time. As a result, people and by extension organizations, find it hard to keep track of these trends and are often lost in a sea of topics. For example, we know that artificial intelligence (ai) is a hot topic in the recent years. However, we are unable to tell its various applications in other domains.

This project aims to use natural language processing techniques to analyse tech articles from online sources, derive topics from these articles, and analyse tech trends to allow easier monitoring of such trends across a period of time. There have been studies and projects to classify news articles into their different topics using the BBC news dataset. However these are based on broad topics (business, entertainment, politics, sport or tech.) which do not drill down into the different subtopics of each domain. This project would allow tech professionals and corporate communications teams to better understand the landscape and quickly detect emerging trends within a tech domain.

Data Source

The data source consisted of tech articles from tech websites such as techradar, wired etc.and mainstream news sites focused on reporting tech news such as reuters, new york times and the guardian etc. Data was retrieved by querying the newsapi with keywords related to the categories of interest namely: ai, cybersecurity, electric vehicles, cloud (computing), blockchain and digitalisation. These keywords acted as a label for validation purposes. The api returned links and information to the articles including a short summary. As the short summary is insufficient to do textual analysis, we used newspaper3k to obtain the full article using the link provided by the api. The dataset includes the columns date, title, source, blurp, text and label.

For the training set, a total of 1200 articles were retrieved and after preprocessing, a step which will be explained later, there were 480 articles remaining with a vocabulary size of 74556. As newsapi only allows users to query articles from a month ago, the dates of the articles ranged from 25 Mar 2022 to 22 Apr 2022.

The test set consisted of 71 articles from 22 Apr 2022 to 27 Apr 2022. This was to mimic a real use case where a company uses the model on a new and recent set of data.

	title	imgurl	date	blurp	url	text	source	
0	What Is Machine Learning, and How Can It Help ...	https://assets.entrepreneur.com/content/3x2/20...	27/03/2022	Artificial intelligence helps create content m...	https://www.entrepreneur.com/article/420562	Opinions expressed by Entrepreneur contributor...	entrepreneur	en c
1	First autonomous X-ray-analyzing AI is cleared...	https://cdn.vox-cdn.com/thumbor/Na1FxfmSilyVIK...	05/04/2022	An X-ray-analyzing tool called ChestLink has a...	https://www.theverge.com/2022/4/5/23011291/ima...	An artificial intelligence tool that reads che...	the verge	ir
2	Australian leagues sign deal to combat online ...	https://www.reuters.com/resizer/MI-ukx_nTJluh9...	03/04/2022	Football clubs in Australia will use artificia...	https://www.reuters.com/lifestyle/sports/austr...	April 3 (Reuters) - Football clubs in Australi...	reuters	
	Does AI get			How different		How different would		

Methodology

This section will cover the methodology used in the preprocessing for the text column. This included: the removal of duplicates from the articles, the grouping of articles with multiple categories together (ie. an article can be both about blockchain and artificial intelligence), the elimination of short articles with less than the removal of punctuation, numbers, and stopwords, using pos tagging using the spacy package to remove intermediate words such as prepositions, adverbs verbs, modal verbs (to ensure that the model is able to pick up the correct keywords), tokenization of the text to form individual words as well as the creation of a bigram model to ensure words such as artificial intelligence will be considered as a single entity compared to separate words.

To ensure words with the same base to be recognized as a single entity, lemmatization was performed. Lemmatization considers the context of the word before converting it to its base form (eg. data to datum, vehicles to vehicle).

The data from the text column was transformed into a document term matrix. Each row of the data would be an article (document), while the features will be the words from the article (Each column represents a word). This forms the corpus to perform topic modelling on.

Topic modelling using latent dirichlet allocation (LDA) (Blei et al., 2003) was used to obtain topics from these articles. The assumptions following LDA are that 1) articles (or documents) contain different topics 2) topics are formed by a mixture of words. As such, words are being generated by a probability distribution of each topic, while articles form the probability distribution of topics. The number of documents will be represented by the variable M, whereas the number of words will be N.

LDA breaks down the document (article) word matrix into two separate matrices, the document topic matrix and the topic word matrix. The former matrix will have a chosen number of topics which form the columns, while the documents form the rows. For the latter, the rows are denoted by topics and the columns will be the words a topic contains.

Following LDA approach, each word will have a latent (related) topic represented by Z. The probability of Z being assigned to a topic will follow θ which is the topic distribution. LDA possesses two controls: α and β. α controls the document topics distribution, whereas β controls for the topic word distribution. A higher α would result in documents having more topics while lower α results in less. For β the higher it is the more words appear in a given topic and vice versa. This project utilized the gensim package with alpha set as 'auto' allowing the algorithm to learn the prior asymmetric prior from the corpus. Similarly beta or 'eta' in gensim, was set to auto to learn the asymmetric prior. Attempting to change either alpha or beta into symmetric or asymmetric resulted in less coherent topics and a larger divergence from the original keyword which was queried.

According to Gan and Qi, the LDA algorithm ranks words based on the probability distribution and use this result as their output. This ranking of words in turn represents the importance of the word to the topic. The results returned by the algorithm will be in the form of words and the given weights for the topics ie. which words represent a given topic:

```
[(0, '0.021*"ev" + 0.017*"car" + 0.015*"battery" + 0.015*"vehicle" + 0.010*"production" + 0.010*"electric" + 0.010*"carbon" + 0.009*"energy" + 0.008*"model" +
0.006*"emission"'), (1, '0.019*"datum" + 0.013*"business" + 0.012*"service" + 0.012*"cloud" + 0.011*"market" + 0.010*"customer" + 0.009*"solution" + 0.008*"pla
tform" + 0.008*"digital" + 0.007*"security"'), (2, '0.023*"security" + 0.016*"vulnerability" + 0.011*"threat" + 0.008*"quest" + 0.007*"code" + 0.007*"memory" +
0.007*"vr" + 0.006*"hardware" + 0.006*"attack" + 0.005*"actor"'), (3, '0.013*"government" + 0.009*"group" + 0.009*"firm" + 0.008*"country" + 0.007*"security" +
0.006*"cybersecurity" + 0.006*"company" + 0.006*"attack" + 0.005*"remittance" + 0.005*"policy"'), (4, '0.048*"blockchain" + 0.034*"game" + 0.012*"player" + 0.0
12*"transaction" + 0.012*"card" + 0.011*"crypto" + 0.011*"cryptocurrency" + 0.011*"nft" + 0.008*"digital" + 0.008*"fund"'), (5, '0.026*"ai" + 0.012*"system" +
0.010*"human" + 0.009*"computer" + 0.008*"error" + 0.008*"quantum" + 0.007*"model" + 0.007*"code" + 0.006*"work" + 0.006*"qubit"'), (6, '0.022*"stock" + 0.020
*"price" + 0.014*"growth" + 0.014*"sale" + 0.013*"revenue" + 0.013*"quarter" + 0.012*"brand" + 0.012*"market" + 0.010*"high" + 0.010*"rating"')]
```

To determine the optimal number of topics for the LDA model, we took a two pronged approach. First, this study trialled a varying number of topics for LDA before settling on be 7 topics after analysing the keywords tagged to each topic and the plot by PyLDAvis to ensure that the topics had a good intertopic distance. Second, the number of topics was decided by selecting the model with the highest coherence score.

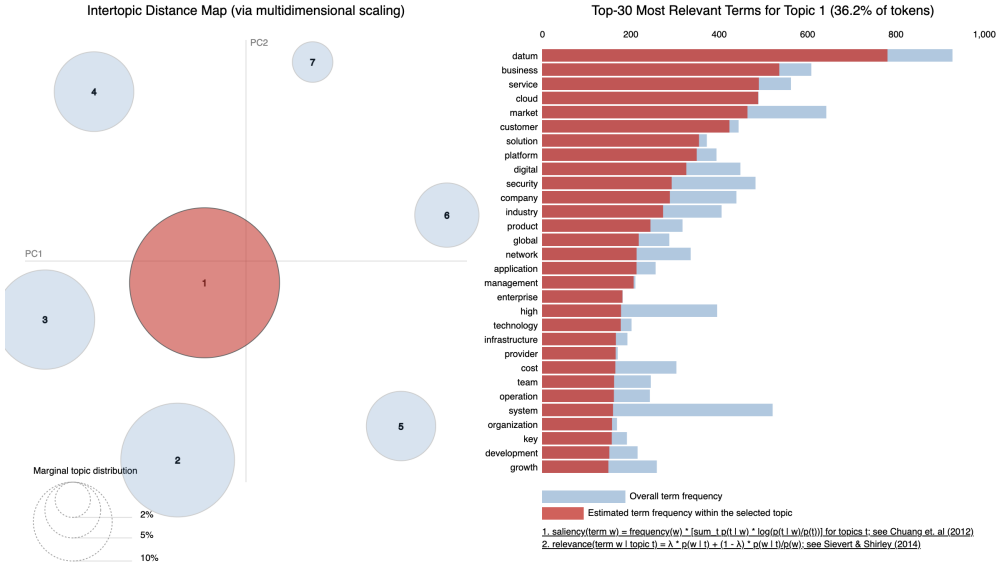
Once the model was trained with the determined number of topics, we obtain the topic distribution for each document and assign the topic to the document if the probability is above the threshold of 0.2.

```
{0: 'ev',
 1: 'business/cloud/digitalisation',
 2: 'cybersecurity/cyberattacks',
 3: 'government/cybersecurity/digitalisation',
 4: 'blockchain/cryptocurrency',
 5: 'ai',
 6: 'market/ev'}
```

Taking into account the words and weights from the given output, human intuition was required to determine a logical "topic" from the weighted tokens. This requires an understanding of the topic covered (for example a combination of ev, energy, vehicle and car could be referring to the topic on Electric Vehicles). This step relied on human judgement as it requires a human to decide the overall topic. Unlike the keyword (categories) which were used to query the api (artificial intelligence, digitalisation etc.), the topics tended to be broader than these categories and usually included a mix of categories. This means some other contextual categories were added for informational purposes. For example, cybersecurity was linked to government, while digitalisation also appeared frequently with cloud. Increasing the number of topics did not result in the segregation of topics. This could mean that the topics selected beforehand were subtopics of a larger topic. As such, the topics assigned to the articles by LDA had multiple categories. Hence, several categories from the original queries had to be merged together.

To ensure that the merging was systematic, the rule applied was as such: if the category of interest (among the 6 original queried keywords) or its synonym appears in the top 30 (usually top 10) it can be included in the current topic. Topic 0 is largely about electric vehicles and the companies selling them. Topic 2 covers cloud platforms and services as well as how business utilize them for digital transformation. Topic 3 is related to cybersecurity issues, ranging from cyber attacks to threats and vulnerabilities, and security issues as well. Topic 4 covers blockchain and crypotocurrency, covering subjects such as bitcoin, ethereum and nfts. Topic 5 encompasses artificial intelligence and its use cases. Topic 6 covers market related issues. This consisted of stock prices, sales, growth, profits, revenue of companies. The category of Electric vehicles was included as it is a hot subject within this topic.

Evaluation and Final Results



The above shows the distribution of the topics with regards to percentage of tokens. From this figure and the previous, we can infer the most popular topic in the corpus in the selected timeframe.

Validation 1

Validation was performed by comparing the topic/ topics assigned, to the original keyword (category) used in querying the newsapi. Due to topics being constituted by multiple keywords, as long as the topic the article was assigned to contained the initial keyword, it was considered as correct. This was a logical step as it was unexpected for articles to have only a single category when articles cover multiple categories or topics to begin with. Using this approach we achieved a 89% accuracy in the training set and a 83% in the test set in predicting the original categories. Reducing the threshold for assigning topics to 0.15 resulted in an increased accuracy of 92.5% and 90% respectively but introduces more labels to the documents.

Validation 2

For the second validation, the top 5 articles from each topic assigned by the algorithm were used to determine whether the article was assigned to the correct topic(s). After reading through the article, one can assess if the topic it was assigned to was correct. As this part will be highly subjective we could set the threshold as the article having at least one correct topic. This means if the LDA topic we gave is "artificial intelligence" and after reading the sampled article and labelling it the same topic, we consider this to be labelled correctly. This was done for both the training set and the test set.

```
0 topic: ev
--------------------------
287
Assigned topic: ev
Assigned topic (multi): ['ev']
Original Topic: ['EV']
text: If you thought it wouldn't be long before Kia turned the Concept EV9 into a production model, you guessed correctly. The automaker has revealed that a ro
ad-ready version of the electric SUV will be available in Europe in 2023. There was no mention of launches in North America or other regions, but it's an SUV —
it may just be a matter of time before you see the EV9 cruising American and Canadian streets.

Kia didn't say what would change in the transition from concept to production. However, we'd expect the badge to cut many of the more exotic features, includin
g the yoke, giant wheels, retractable roof rails and lounge-like seating modes. We wouldn't be surprised if Kia kept the 27-inch display and even the hood-moun
ted solar panel, though.

The production EV9 might also preserve the claimed specs. The concept promised up to 300 miles of range and 350kW fast charging that could take it from a 10 pe
rcent charge to 80 percent in 30 minutes. Kia also recently detailed autonomous "Automode" technology for the EV9 that can take over from the driver on the hig
hway.

There are still important unknowns like pricing. Even so, the EV9 could be one of Kia's most important all-electric vehicles to date, at least in some areas. W
hile the EV6 has been well-received so far, some markets (particularly North America) skew heavily toward crossovers and SUVs. The EV9 could help Kia take on c
ompetitors like the Tesla Model Y and Volkswagen ID.4, not to mention reel in buyers who haven't been thrilled by the Niro EV.
```

From the above, it can be inferred that the topic of EVs is correct as the paragraph is referring to the Kia's model EV9 and mentions competitors such as Tesla.

```
Assigned topic: cybersecurity/cyberattacks
Assigned topic (multi): ['cloud/digitalisation', 'cybersecurity/cyberattacks']
Original Topic: ['cloud']
text: Cloud computing and virtualization technology firm VMWare on Thursday rolled out an update to resolve a critical security flaw in its Cloud Director prod
uct that could be weaponized to launch remote code execution attacks.

The issue, assigned the identifier CVE-2022-22966, has a CVSS score of 9.1 out of a maximum of 10. VMware credited security researcher Jari Jääskelä with repor
ting the flaw.

"An authenticated, high privileged malicious actor with network access to the VMware Cloud Director tenant or provider may be able to exploit a remote code exe
cution vulnerability to gain access to the server," VMware said in an advisory.

VMware Cloud Director, formerly known as vCloud Director, is used by many well-known cloud providers to operate and manage their cloud infrastructures and gain
visibility into datacenters across sites and geographies.

The vulnerability could, in other words, end up allowing attackers to gain access to sensitive data and take over private clouds within an entire infrastructur
e.
```

Another example would be this article which talks about cloud computing with respect to cybersecurity where VMware rolled out a patch to resolve a security flaw in its cloud product. Hence, after evaluation, we believe that the topic assigned to the article is appropriate and it explains more compared to its original topic of cloud computing.

```
----------------------
1 topic: business/cloud/digitalisation
-------------------------
414
Assigned topic: business/cloud/digitalisation
Assigned topic (multi): ['business/cloud/digitalisation']
Original Topic: ['cybersecurity']
text: Opinions expressed by Entrepreneur contributors are their own.

For the better part of the last decade, the world has experienced dramatic changes. These changes have had huge impacts on the way businesses operate. As a res
ult, many organizations have been forced to shift to a digital model. The number of businesses succeeding in digital transformations continues to increase ever
y day.

Digital transformation means using modern technologies to improve business operations, systems, processes and customer experience. The transformation lowers th
e operating expenses of many firms. The workers' performance and productivity also increases, boosting the company's profitability.

For every modern business to enjoy success in the market, it needs to accelerate its digital transformation. Here is an overview of five practical strategies t
hat we have seen to be the top focus for companies in 2022 to help firms speed up their digital transformation efforts.

1. Finding the right technology

Digital transformation is not only about the current software in the company. Finding the right technology is essential for a faster transition. Hiring managed
IT services from experience providers is one way of avoiding pitfalls during the transition.
```

Using the second validation approach, most of the articles contained their originally assigned categories. For the above article, the algorithm correctly classified the article under business/cloud/digitalisation (which covers businesses,services) organizations while the original category of cybersecurity was not representative. The article explains methods to speed up digital transformation for businesses, whereas cybersecurity was mentioned as a sidenote in the article.

```
5 topic: ai
-------------------------
10
Assigned topic: ai
Assigned topic (multi): ['cloud/digitalisation', 'ai']
Original Topic: ['ai']
text: We are excited to bring Transform 2022 back in-person July 19 and virtually July 20 - 28. Join AI and data leaders for insightful talks and exciting netw
orking opportunities. Register today!

Arize, a maker of artificial intelligence (AI) observability tools, has introduced Bias Tracing, a new tool for identifying the root cause of bias in machine l
earning (ML) pipelines. This can help teams prioritize and adjustments and address issues either in the data or the algorithm itself.

Enterprises have long used observability and distributed tracing to improve applications performance, troubleshoot bugs and identify security vulnerabilities.
Arize is part of a small cadre of companies adapting these techniques to enhance AI monitoring.

Observability analyzes data logs to monitor complex infrastructure at scale. Tracing reassembles a digital twin representing the application logic and data flo
w for complex applications. The new bias tracing applies similar techniques to create a map of AI processing flows spanning data sources, feature engineering,
training and deployment. When bias is detected, this can help data managers, scientists and engineers root out and rectify the root cause of the problem.

"This type of analysis is incredibly powerful in areas like healthcare or finance given the real world implications in terms of health outcomes or lending deci
sions," said, Aparna Dhinakaran, Arize cofounder and chief product officer.
```

```
32
Assigned topic: blockchain/cryptocurrency
Assigned topic (multi): ['business/cloud/digitalisation', 'blockchain/cryptocurrency']
Original Topic: ['blockchain']
text: Ark Invest Chief Executive Officer Cathie Wood has never shied away from bold predictions. In 2018, she put a price target on Tesla that implied a $672 b
illion market cap. Of course, Tesla has since exceeded that valuation by leaps and bounds -- but at the time, the company was worth just $56 billion.

Wood is also a well-known crypto bull. In fact, a recent report from Ark Invest suggests that Ethereum (ETH -0.42%) could achieve a valuation of more than $20
trillion in the next 10 years. That implies 5,400% upside from its current price. Given Wood's bullish outlook, let's take a closer look at this cryptocurrenc
y.

Here's what you should know.

The first programmable blockchain
```

The same validation was done for the test set. We observe that the topic model was able to predict the correct topics for the test set as shown from the above examples.

```
Assigned topic: government/cybersecurity/digitalisation
Assigned topic (multi): ['business/cloud/digitalisation', 'government/cybersecurity/digitalisation']
Original Topic: ['ai']
Correct?: 0
text: We are excited to bring Transform 2022 back in-person July 19 and virtually July 20 - 28. Join AI and data leaders for insightful talks and exciting netw
orking opportunities. Register today!

Strider, a platform that helps companies, governments, and research institutions protect intellectual property (IP), talent, and supply chains from nation-stat
e threats, has raised $45 million in a series B round of funding.

Founded in 2019, Salt Lake City-based Strider pitches two core products. Strider Risk Intelligence leverages disparate datasets spanning industrial policy docu
ments, international patent data, among other publicly-available sources, to help companies identify which of their core technologies may be most at-risk from
nation-state actors.

Feeding into this are "risk signals," which meshes thousands of primary data sources with a "proprietary risk methodology" to spot high-risk activities, and gi
ve organizations data to act upon.

Strider: Risk intelligence
```

```
Assigned topic: cybersecurity/cyberattacks
Assigned topic (multi): ['business/cloud/digitalisation', 'government/cybersecurity/digitalisation']
Original Topic: ['ai']
Correct?: 0
text: We are excited to bring Transform 2022 back in-person July 19 and virtually July 20 - 28. Join AI and data leaders for insightful talks and exciting netw
orking opportunities. Register today!

CrowdStrike has unveiled new capabilities for its adversary-focused cloud-native application protection platform (CNAPP). These new capabilities shorten the ti
me it takes to respond to threats in cloud environments and workloads by accelerating threat hunting.

CrowdStrike specializes in cloud-delivered endpoint protection, cloud workloads identity and data. CrowdStrike Security Cloud and world-class AI operate on the
CrowdStrike Falcon platform. This platform employs real-time attack indicators, threat intelligence, developing adversary trade craft and enriched telemetry fr
om across the enterprise, to enable hyper-accurate detections, automated protection and remediation, elite threat hunting and prioritized visibility of vulnera
bilities.
```

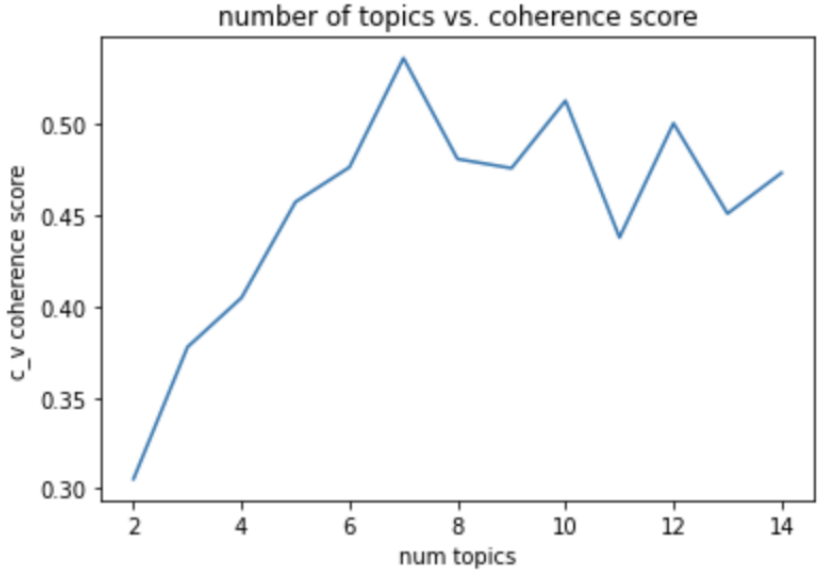
However, there are instances where the model was not able to predict the original category of the article. This was because the article contained more information related to the assigned topic (by the lda model) and only a few words containing the original category (such as ai). The above articles demonstrate the following: the article is regarding a cloud application with cybersecurity capabilities, yet the original category was ai.

Validation 3

A good lda model would usually be interpretable (Roder et al., 2015) and a way to measure this is using coherence score. Coherence measures the similarity of words within a topic generated by the model. The more similar words are within a topic, the higher the coherence score and the better the model. A way to measure coherence is based on Normalized Pointwise Mutual Information (NPMI). Elements of vectors defined using NPMI were found to be highly correlated with human topic coherence ratings which are considered the gold standard (Roder et al., 2015). A context vector of word w takes the occurrence of 5 surrounding words (before and after) in relation to w forming a vector. This means that in a word vector, the jth element of w_i will have the formula:

$$NPMI(w_i, w_j)^\gamma = (\frac{log\frac{P(w_i,w_j)+\epsilon}{P(w_i)P(w_j)}}{-log(P(w_i, w_j) + \epsilon}})^\gamma$$

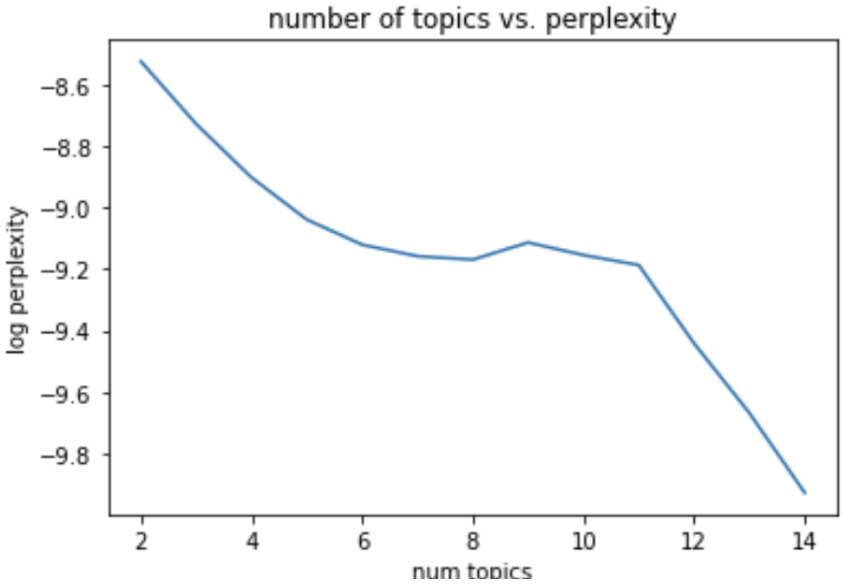
From the plot of the coherence score below,we observe that having 7 topics result in the highest coherence score of 0.53. As this score was the best, and the topics yielded were understandable, the finalised model contained 7 topics.



Validation 4

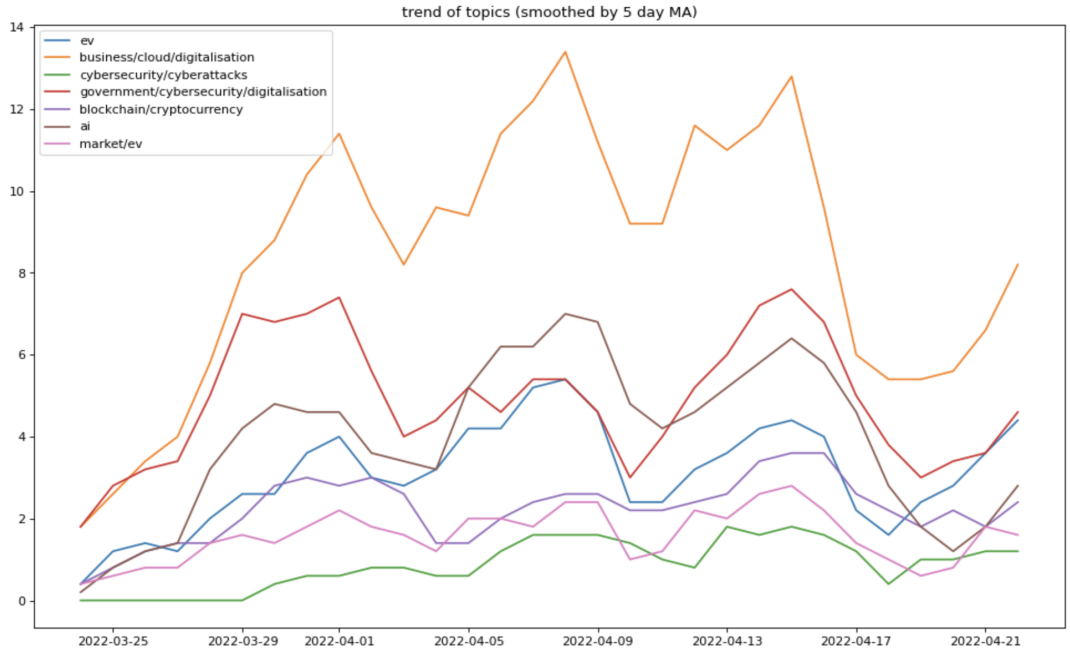
$$PP(W) = P(w_1, w_1...w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1, w_1...w_N)}}$$

The perplexity of the model is a metric which reflects how well a model will perform on a new set of data. According Jurafsky and Martin, the perplexity of a model on the test set is the inverse probability on the test set. It is known as an intrinsic measure for evaluating a natural language model. The higher the perplexity the better the model performs on unseen data. For this study we compared the perplexity of the training set to the test set. In gensim, the model outputs a log of perplexity causing the values to be negative. This means that a less negative perplexity will be better than a more negative one. The perplexity on the train set was -7.87 while the test set was -9.15. The difference is not too large and it is normal for a models. A plot of perplexity versus the number of topics (on the test set) shows that the bottom the curve is about 7 or 8 topics, this lends support to the current model.



Results

The results of this project is the assignment of these topics to articles. With this we can track the trending topics over a period of time. This could be in the form of a simple visualization (multi line charts with the count of topics on the y axis and date on the x axis.



From the line chart of the topic 3 topics, we see that business/cloud/digitalisation is the most dominant topic and we can view a sharp spike from 2 Apr 2022. government/cybersecurity/digitalisation follows the trend of the former but has less of a spike. Moreover, while both topics started with the same number, the former became more popular over time as compared to the latter.

A second interesting point would be when a topic becomes more popular than another. For example, in the period of 5 Apr to 10 Apr, artificial intelligence superceded government/cybersecurity/digitalisation as the second hottest topic. Therefore, using this visualisation, a company could monitor technology trends and it could serve as a guide for their overall strategy.

Potential Limitations

A possible limitation of the approach would be that topics might shift over time and having a fixed number of topics could result in the model being unable to pick out newer topics. To counter this issue, the number of topics chosen has to be tweaked over time to account for different topics emerging.

Secondly, the queried keywords might not necessarily reflect the all the topics discussed within an article. Articles tend to be interdiscplinary and cover a range of topics. There have been instances where the topic model was able to pick up the topic better than the original categories. Therefore, the first validation method was limited by the accuracy of the original categories.

This leads to the next point that as LDA is an unclassified machine learning algorithm, the results might not fit neatly into the original categories. Take for example the output of weighted tokens per topic, we see digitalisation present in multiple topics, instead of forming its own topic. Hence, the the second validation approach is required to ensure that the topics match the assigned categories. This in turn requires human intuition and effort to evaluate the articles and could be difficult to implement with datasets with more than 20 topics.

Lastly, as the number of articles is small, with only a months worth of data, the topic model is affected by topics which are recent. This issue can be solved by obtaining articles across a longer

period to ensure that topics will be more stable across time. Additionally, most projects run topic models on a larger corpus of about 100,000 documents. Therefore, while this project has demonstrated satisfying results with a small corpus, a further improvement would be to collect more data and train the model on a larger corpus.

Conclusion

This project aimed to find the most trending topics in technology over a time period and was able to demonstrate the ability to do so with the use of LDA. With the knowledge of these trends, organisations have a better understanding of the applications in different areas of tech and are better able to make decisions regarding long term strategies. As the dataset is limited in this study, we believe that with a larger dataset, the topic model will yield more stable topics allowing organizations to keep better track of the shifting trends.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Gan, J. & Qi, Y. (2021). Selection of the optimal number of topics for LDA Topic Model—taking patent policy analysis as an example. Entropy, 23(10), 1301. https://doi.org/10.3390/e23101301

Jurafsky , D. & Martin J. H., (2021). Speech and Language Processing (pp. 8–9). Draft.

Röder, M., Both, A., &; Hinneburg, A. (2015). Exploring the space of topic coherence measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. https://doi.org/10.1145/2684822.2685324