**Topic modelling with LDA**

Andreas Rouvalis

**1. Introduction: Topic modelling**

The project centers around topic modeling, a statistical modeling technique designed to identify recurring themes and patterns within a specified dataset. As a form of unsupervised learning, it falls within the broader realm of text mining. Semantic structures are harnessed to capture shared abstract topics across a collection of documents and ultimately understand and organize a corpus. Essentially, emerging topics are clusters of co-occurring words created using a probabilistic approach and through which themes are discovered.

Several algorithms have been proposed, including Non-negative Matrix Factorization, Structural Topic models, Latent Semantic Analysis, and Latent Dirichlet Allocation (LDA). Among these, LDA introduced by Blei, Ng and Jordan (2003), stands out as the most commonly used algorithm and is the one employed for the purposes of this project.

**2. Latent Dirichlet Allocation**

Latent Dirichlet Allocation (LDA) is a Bayesian generative statistical model that considers a given corpus as a collection of documents each consisting of a sequence of words. This can be represented 2-dimensional matrix of words *WNxM*, where *M* is the number of documents and *N* is the maximum number of words per document.

In the LDA generative process, documents are not generated in the linguistic sense of structured set of sentences but rather as bags of words. The generation of each document *d* in the corpus follows the procedure depicted in Figure 1 using plate notation. The outer plate represents the document, and the inner plate indicates the repetitive selection of topics and words within each document.

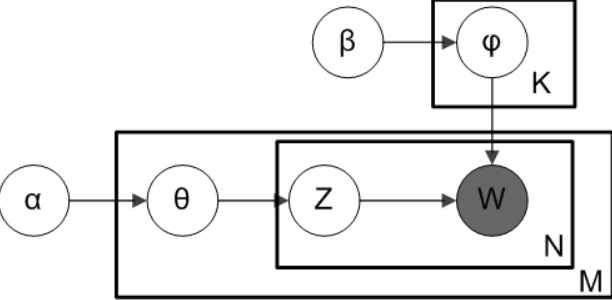


Figure 1 Generative process

This can be described as a two-step process:

1. *d ~ Dir(α)*: Generate the document-topic distribution for each document.
2. For each of the Nd words in ***d***:
   1. Choose a topic *k* ~ Multinomial(*d*).
   2. Choose a word *wn* from *p(wn | k, β)* = *φk  ~ Dir(β)*

The parameters α and β determine the document- topics distributions (*Dir(α)*) and topic-words distributions (*Dir(β)*) respectively. These parameters determine whether the distributions are uniform (value =1), concentrated (value > 1) or sparse (value < 1). By choosing values lower than 1 the probabilities will follow Dirichlet distributions.

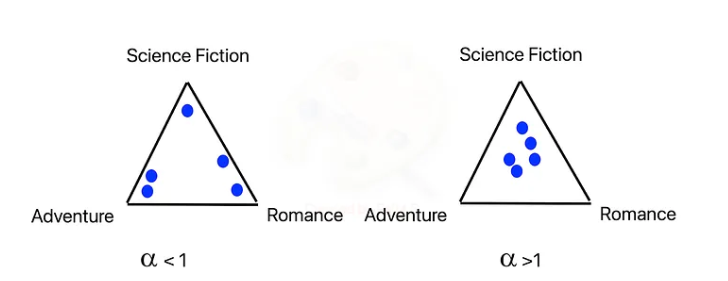
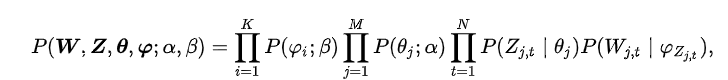


Figure 2 The effect of α

Dirichlet distributions imply that each document consists of a small number of dominant topics and each topic is characterized by a small number of words. However, when the parameter *α* is set to a higher value, it results in a model with numerous dominant topics (see figure 2). Similarly, a higher value for *β* places less emphasis on each topic being composed of only a few predominant words. Both d and *φk*, chosen from such distributions are latent variables to the model, in contrast to the words which are observable variables.

The total probability of a generated corpus is a joint probability given the parameters is:

 (1)

Here *w*, *z* and *θ* are vectors of the variables. *z* is produced during the generative process which assigns each generated word a specific latent topic. Similarly, *w* is generated, representing the vector of unique words.

The goal of topic modeling, however, is not to generate the corpus but to infer the hidden topics, essentially reversing the generative process described above. In technical terms, we need to solve the following equation:

(2)

Here the document-topic distribution (*θ*), topic-word distribution (*φ*), word-topic assignment (*z*) is inferred given the parameters *α*, *β* and corpus *w*. This distribution is intractable to compute as the normalization factor cannot be calculated exactly. Since the latent variables *θ* and *φ* can be intergraded out during the generative process, we can instead compute *P(z| w; α, β)* using an inference approximation technique such as Gibbs sampling.

Gibbs sampling, a Markov Chain Monte Carlo (MCMC) algorithm, is a technique used for approximating inference in probabilistic models. In the context of topic modeling and LDA, Gibbs sampling can be employed to estimate the posterior distribution of latent variables. Various forms of Gibbs sampling are available, and for this project, I employed the technique known as collapsed Gibbs sampling. Collapsed Gibbs sampling simplifies the inference process by marginalizing out some of the latent variables, resulting in a more computationally efficient approach. It approximates the probability topic *k* is assigned to a word *wi*through the following iterative process (see also pseudocode in figure 3 taken from Darling, 2011).

1. **Initialization**: Randomly initialize *z* which represents the topic assignment of words.
2. **Iterative Sampling:**
3. Sample a new topic for each word from the distribution *p(Zn,m = k|·)* which is inferred from the current *Z* excluding the current assignment for that specific word using the following equation:

(3)

1. Update *Z* for each word in the corpus based on the sampled values.
2. **Repeat**: The iterative sampling process is repeated for a predetermined number of iterations, allowing the Markov Chain to converge to the posterior distribution *P(z|w;α, β)*.

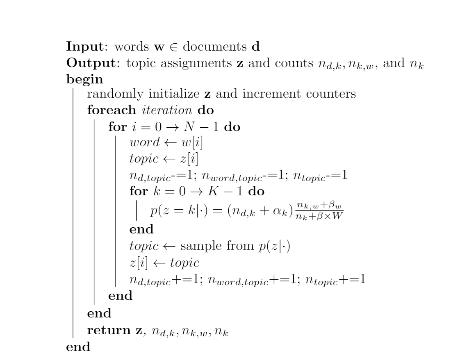


Figure 3 Pseudocode for Gibbs Sampling

3.**The LOCO corpus**

The collection of documents was sourced from the Language of Conspiracy (LOCO) corpus, comprising 88 million tokens. This dataset contains both conspiracy theory documents (N= 23,937) and mainstream (non-conspiracy) documents (N=72,806). Conspiracy theories texts are defined as texts containing narratives that explain significant social events as schemes orchestrated by powerful individuals or groups (Douglas et al, 2019).

Documents were extracted from a set of 150 websites. Websites containing conspiracy theory documents were identified through a website dedicated to fact-checking online resources and flag websites that publish unverified information about conspiracy theories[[1]](#footnote-1). In contrast, mainstream document domains were identified via Google search.

Specific documents from the websites were identified using ‘*seeds*’, a set of keywords and key phrases encompassing a broad spectrum of conspiracy theories. The final set of seeds was based on two conspiracy theory surveys (Jensen 2013; Douglas and Sutton, 2011), with an additional 20 seeds introduced by the corpus creators to enhance dataset diversity. This was necessary as surveys did not include newly emerged conspiracy theories (e.g. COVID-19, pizzagate). It should be noted that the seeds overlap (e.g., "new world order" and "NWO," "climate change" and "global warming") to account for different spellings.

For this project, I focused on documents from the dataset that pertain to 5G technology, both conspiracy theory and mainstream documents. Introduced globally in 2019 as a successor to 4G, 5G technology provides higher capacity, increased bandwidth, and lower latency, enhancing communication experiences. However, its emergence also triggered a multitude of conspiracy theories worldwide, rapidly, ironically, across online platforms.

**4. The Project**

The goal of the project is to identify prevalent topics from the collection of 5G conspiracy theory documents (N= 702) as well as the collection of mainstream documents (N= 1664), discussing 5G technology and compare the topics identified in these two sub-corpora.

LDA was implemented in two ways. Firstly, leveraging the capabilities of Gensim, a Python library tailored for facilitating topic modeling tasks. Secondly, I implemented the Latent Dirichlet Allocation (LDA) algorithm without reliance on any pre-existing topic modelling libraries.

**5. Discussion**

Determining the optimal number of topics stands as the foremost important task in topic modeling. Given the potential computational cost of executing LDA from scratch, especially when handling sizeable corpora, leveraging the Gensim implementation facilitated multiple runs of the model to identify the appropriate number of topics. Commencing with an initial high number of topics, I systematically reduced this count until achieving a set of topics that were both coherent and interpretable- to the extent that was possible. Subsequently, I executed my custom LDA implementation. Different hyperparameters were tried (e.g. alpha = 0.005, 0.02,0.001, beta = 0.05, 0.2, 0.1, passes = 300, 500, 600, 800). The hyperparameters that gave rise to the clearest topics were: alpha = 1/number of topics, beta =1/number of topics. Notably, to yield cohesive topics using my custom LDA implementation, a higher number of passes were required (600 passes for the conspiracy corpus and 500 passes for the mainstream corpus) compared to those required for the library-assisted implementations which gave rise to coherent topics after 300 passes in both corpora.

Four distinct topics were identified within the conspiracy theories corpus. The table displays the words output by both models, along with manually added titles for further clarification.

|  |
| --- |
| The health effects of 5G technology   * "radiation", "study", "exposure", "health", "research", "frequency" (Topic 1-Gensim) * "radiation", "study", "exposure", "health", "research", "body" (Topic 1-My LDA model) |
| The connection between 5G and the COVID-19 pandemic   * "virus", "vaccine", "coronavirus", "covid19", "disease", "child"(Topic 2-Gensim) * "virus", "vaccine", "coronavirus", "covid19", "death", "world" (Topic 2- My LDA model) |
| 5G, global politics and economy   * "trump", "economy", "money", "market", "company", "world"' (Topic 3-Gensim) * "trump", "money", "economy", "dollar", "world", "debt" (Topic 4- My LDA model) |
| Topic 4a: 5G, Power, and Humanity's Control   * "world", "control", "energy", "power", "humanity", "life"(Topic 4-Gensim)   Topic 4b   * "world", "company", "project, "datum", "power", "information"(Topic 3-My LDA model) |

*Table 1: Topics emerging from the conspiracy corpus*

The initial two topics are coherent and distinctly encapsulate themes within the narrative of 5G conspiracy theories. Topic 1 delves into the perceived health hazards of 5G technology such as the belief among conspiracy theorists that 5G technology is used to expose humans to harmful radiation. Topic 2 focuses on the connection between COVID-19 and 5G. Conspiracy theorists claim that 5G technology transmits the virus.

In contrast, Topic 3 encompasses broader terms associated with global politics and the economy, with notable references to the former USA president, a recurring figure in conspiracy theory narratives. Lastly, the fourth topic is the most conceptually vague one and seems to correspond to the re-occurring theme of conspiracy theory narratives that there is a scheme whose ultimate aim is global control. However, the fourth topic generated by my LDA implementation, while containing some words that occur in the fourth topic of the Gensim library, appears even less coherent and significantly noisier.

Turning to the mainstream corpus, 3 topics were identified using the Gensim package:

|  |
| --- |
| Technological Advancements in 5G from a user perspective   * “device”, “datum”, “user”, “communication”, “application”, “phone”, “speed” (Topic 1) |
| Conspiracy theories about 5G   * “health”, “coronavirus”, “exposure”, “study”, “theory”, “covid19”, “virus”’ (Topic 2) |
| 5G and business   * “company”, “market”, “plan”, “tmobile”, “customer”, “infrastructure”, “decision”’ (Topic 3) |

*Table 2: Topics emerging from the mainstream corpus (Gensim)*

Topic 1 primarily revolves around the user-facing aspects of 5G technology, comprising terms related to 5G-enabled devices (e.g., phones), user experience (e.g., speed, communication), and data usage (e.g., datum). On the other hand, Topic 2 delves into discussions surrounding conspiracy theories linked to 5G. While this corpus predominantly consists of mainstream texts that do not explicitly discuss conspiracy theories, references to such theories are inevitably present, often to refute or debunk them. Lastly, Topic 3 centers on the business implications of 5G technology, covering aspects such as market dynamics, strategic planning, and infrastructure development.

In contrast, the topics identified from my custom LDA implementation in the mainstream corpus exhibited less coherence and alignment with the results from Gensim.

|  |
| --- |
| 5G, market and user   * "company", "device", "datum", "market", "phone", "industry", "infostructure" |
| Conspiracy theories about 5G   * "coronavirus", "virus", "group", "covid19", "theory", "health", "claim" |
| Noisy topic   * "food", "cell", "study", "user", "drug", "communication", "energy" |

T*able 3: Topics emerging from the mainstream corpus (My LDA implementation)*

The first topic from my implementation correspond to Topic 3 from the Gensim results, consisting of words about the business impact of 5G, while also containing terms related to topic 1 (*datum, phone*). Topic 2 is also about COVID-19 and its supposed relation to 5G technologies. In contrast, Topic 3 from my implementation appears to be a noisy topic, containing terms such as "food," "cell," and "drug" that lack clear thematic cohesion. Attempts to alleviate this noise by adjusting the number of topics proved challenging, as reducing the count compromised the coherence of other topics, while increasing it introduced more noisy topics.

The less coherent nature of some topics underscores the challenge of assigning clear, well-specified labels, as highlighted by the inclusion of terms like “*energy*” (topic 4 -conspiracy theory corpus) that appear conceptually disjointed from the rest. This underscores LDA's reliance on word co-occurrence patterns, which may inadvertently include noise or document-specific peculiarities.

It is worth noting that a topic regarding COVID-19 and 5G emerges in both corpora. Topic modeling algorithms generate collections of words based on patterns in the corpus, but they do not provide explanations for the meaning of these collections. LDA operates on a bag-of-words model, considering individual tokens in isolation without accounting for their relationships with other words. Consequently, LDA cannot capture nuanced information as it remains oblivious to the context of words or grammar. While topic modeling algorithms excel at identifying clusters of co-occurring words, they lack the ability to understand the context in which these words occur or assess how topics conceptually relate to one another. In the conspiracy theory corpus, we would anticipate a support of the belief linking COVID-19 with 5G, whereas mainstream documents likely discuss this narrative with the intention of debunking it which is hinted by the word “*claim”*. In other words, while the topics are very similar at the word level, the stance of the documents from which the topics emerge differ significantly. This further highlight that topic interpretation is heavily reliant on the nature of the data and human post-hoc judgment remains essential for contextualizing and interpreting topics accurately.

**Literature**

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

Darling, W.M. (2011). A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling.

Douglas, K. M., & Sutton, R. M. (2011). Does it take one to know one? Endorsement of conspiracy theories is influenced by personal willingness to conspire. *British Journal of Social Psychology*, *50(3)*, 544–552.

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). *Understanding Conspiracy Theories. Political Psychology, 40(S1),* 3–35. https://doi.org/10.1111/pops.12568.

Jensen, T. (2013). *Democrats and Republicans differ on conspiracy theory beliefs. Public Policy Polling.* http://www.publicpolicypolling.com/polls/democrats-and-republicans-differ-on-conspiracy-theorybeliefs.

1. <https://mediabiasfactcheck.com/conspiracy/> [↑](#footnote-ref-1)