Measuring central bank uncertainty: A computational linguistics approach

Péter Horváth

June 04, 2024

Abstract

This is the abstract.

1 Introduction

Motiváció/literature szinten bizonytalan vagyok, hogy mihez kéne tartozni. Leginkább NLP? Uncertainty? Forward Guidance (proxy)? Identifikáció?

A results preview-nak pedig empírikus eredmények híjján még nem álltam neki.

2 Building the Central Bank Uncertainty Indices

Measuring thematic uncertainty within FOMC publications can be grouped into two stages. Firstly, in the spirit of Baker et al. (2016) and Caldara and Iacoviello (2022), paragraphs on the FOMC releases can be tagged if they contain the word stem "uncertain." Calculating the fraction of tagged paragraphs could in theory yield an uncertainty index, however, uncovering the underlying topic associated with uncertainty. For this purpose, one needs to identify common themes that are discussed within these documents. This can be done using Natural Language Processing (NLP) techniques. The most widely used for this purpose is the Latent Dirichlet Allocation (LDA) of Blei et al. (2003).

This process is carried out on the FOMC transcripts, as these provide the most complete account of what is discussed during the meetings. With the estimated LDA, it is possible to retrieve the posterior topic distributions within the transcripts. Additionally, the LDA being trained on transcripts does not constrain the further use of the model, as such posterior topic distributions can be estimated from other releases, such as Meeting Minutes, Statements, or Beige- and Tealbooks.

Transcripts already have seen some use in the literature. A well known example is Romer and Romer (2023) (and preceding works) reading through each transcript to narratively identify monetary shocks, where the FED intentionally raised rates from the equilibrium in order to combat inflation by depressing the economy. Methodologically closer to this paper, Hansen et al. (2018) have used the FOMC transcripts to identify topics in the FOMC transcripts with and LDA model. For this reason, I closely follow their methods, however, some modifications are made where I felt it necessary.

2.1 FOMC releases

Two websites host the FOMC publications, https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm hosts documents released in the latest 5 years, and https://www.federalreserve.gov/monetarypolicy/fomc_historical.htm hosts documents from all preceding years dating back to 1939. The majority of the releases can be grouped into 5 distinct types, these are plotted in Figure 1

- Transcripts[1]: Complete, word-for-word accounting of what is discussed during the FOMC meetings.
- Memorandum of Discussions (Memo): Similar to transcripts, these provide a detailed accounting of meetings discussions, written from a third person perspective.

[1] Conference calls are also labelled as Transcripts on the FOMC historical website, however, due to their sporadic nature and content, these are ignored for the purposes of this paper.

- Minutes: These include the modern FOMC minutes, Historical Minutes and Records of Policy actions. Although these may vary in length from decade to decade, their purpose is to give a medium-length summary of discussions and decisions made during a meeting of the FOMC.
- Statements: Short (1-2 paragraph) public statements that summarize policy decisions.
- ColorBooks: An !(unprofessional)! umbrella term for Red, Green, Blue, Teal and Beigebooks, which provide information and analysis on economic conditions and policy alternatives.

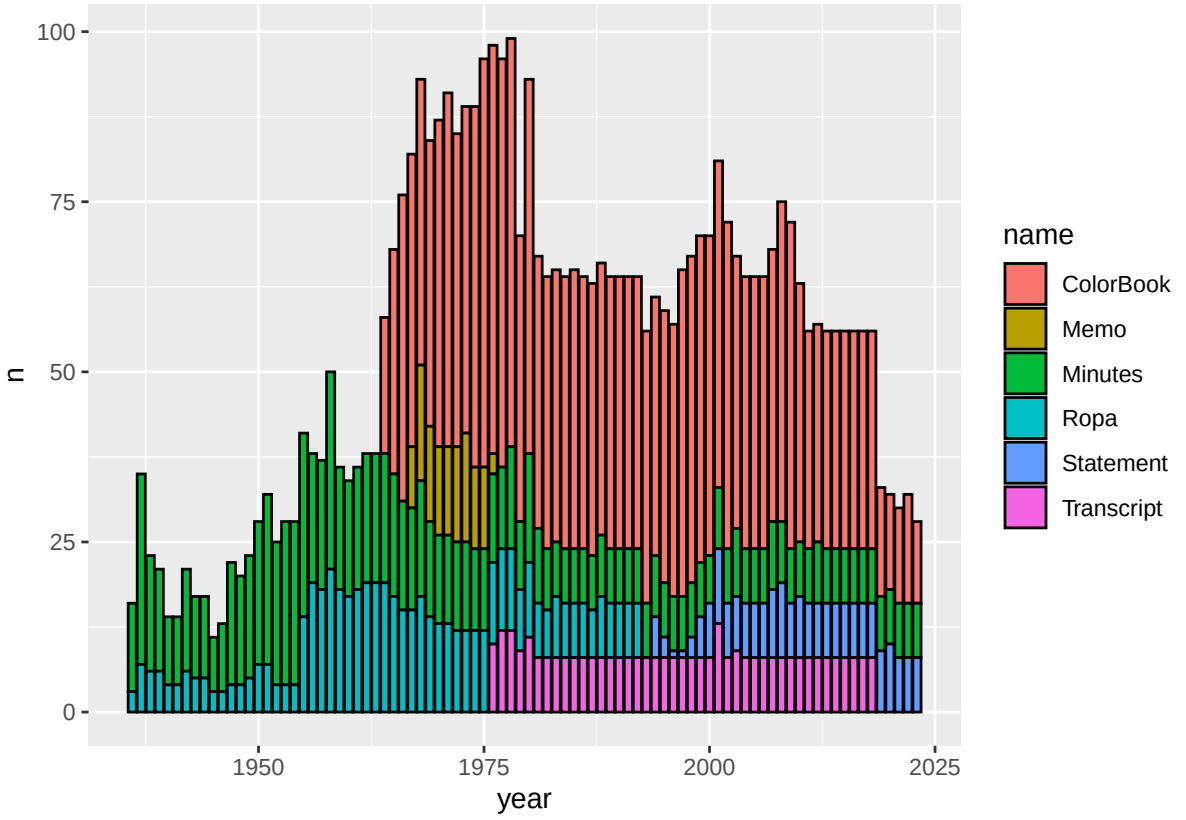

Figure 1: FOMC releases by publication type annually.

Transcripts, available from to as of writing, make up a relatively small proportion compared to other releases, with a total of being available. However, the Transcripts have the unique advantage of giving a full accounting of the discussions during a meeting of the FOMC, as such, these documents are best suited to gauge the proportion of topics based on their importance. From a modelling perspective, while there might only be a handful of transcripts available, each one is close to 200 pages, which chopped into paragraphs, provides a sufficiently large data pool for the purposes of training an LDA model.

2.2 Preprocesing and Vocabulary selection

Textual data is by nature not well suited for quantitative analysis, and as such requires multiple steps to transform it from raw text into a matrix representation suitable for statistical analysis, called a document-term matrix (DTM). A DTM contains rows and columns equal to the number of documents and unique terms found in the text corpus respectively. I chose one paragraph as a unit of measurement for one

document, as in the case of Transcripts it usually refers to an individual interjection, and as such I can expect a similar number of documents as in Hansen et al. (2018). Additionally, in more general terms, a paragraph can be taken as a piece of text covering predominantly one coherent topic, which is ideal for fitting an LDA model.

The data is downloaded from the FOMC calendars and historical websites in PDF[2] form using a web-scraper, and the raw text is mined from these PDFs. The headers and footers are removed using regular expressions[3], and split into paragraphs. Every non-empty paragraph is considered as a document of the corpus.

Next documents are then part-of-speech tagged using the UDPIPE parser of Wijffels (2021) and tokenized into uni-, bi- and trigrams (one, two and three word tokens.) As described in Justeson and Katz (1995), specific word patterns[4] are kept that likely correspond to distinct word collocations. Here I slightly deviate from the methods described in Hansen et al. (2018) by adding single-word tokens. I opted to their additions as the correct accounting of the frequency of unigrams could help better pinpoint each topic and as such may give better interpretability to the topics.

As the last bit of the preprocessing phase, each word of each token is converted to lowercase and stopwords[5] are removed. Each word of each token is then stemmed to its linguistic root, so that words with the same meaning conjugated differently are accounted for in the same feature (column) of the DTM.

In order to streamline the vocabulary, and as such reduce the sparsity of the DTM, I follow Blei and Lafferty (2009) and calculate topic-frequency (TF), inverse-document-frequency (IDF), TF-IDF scores defined as

$$
\begin{aligned}
tf_v &= 1 + log(n_v) \\
idf_v &= log(\tfrac{D}{D_v}) \\
tfidf_v &= tf_v \times idf_v
\end{aligned}
$$

where n_v is the absolute frequency of term v, in the corpus, D_v is the number of documents word v appears in and D is the total number of documents in the corpus. The idea behind the TF-IDF score is the importance of a term is increasing in TF, however, a term present in a large proportion of document should be punished, as its added information to the corpus is likely lower.

Here, I make another important deviation from Hansen et al. (2018). Firstly, I do not pre-define a specific absolute term frequency threshold below which terms are discarded. This is done because in the definitions of TF, IDF and TF-IDF, each score is independent of the total number of words in the corpus. Secondly, I do not discard the lowest ranking words in terms of TF-IDF, instead I discard terms with extreme high IDF values. From the definitions of these scores, this seems counter-intuitive, however I found that

[2] If available, the text was retrieved in HTML form, for a more seamless splitting into paragraphs. HTML sources were in the vast majority only available for modern Minutes and Statements, other publications are fully retrieved as PDFs.

[3]

[4] For bigrams this includes: adjective-noun and noun-noun; for trigrams this includes: adjective-adjective-noun, adjective-noun-noun, noun-adjective-noun, noun-noun-noun and noun-preposition-noun. Additionally only noun unigrams are retained.

[5] Words such as "the," "that," "or," "of."

excluding low TF-IDF scoring terms eliminates some staple words in terms of economic concept, such as "growth," "labor market" or "feder fund rate" due to their low IDF scores. Additionally, terms with extremely high IDF scores mean that the term is present in only an extremely small proportion of documents. Hansen et al. (2018) finds that discarding low TF-IDF score terms primarily removes terms that are low frequency and only present in a small proportion of topics. To put it more concisely, filtering terms based on the IDF score discards extremely sparse terms, while retaining terms that most likely coincide with key economic concepts. I set three IDF cutoff values at 7.6 for unigrams, 8.3 for bigrams and 9 for trigrams[6].

Note to self: Itt érdemes lenne a Hansen et al. (2018) cikkhez hasonlóan egy táblázat a corpus leíró statisztikáiíról az egyes lépések folyamán.

2.3 LDA topic model

Developed by Blei et al. (2003), LDA is a probabilistic model that builds on K-means clustering algorithms by relaxing their binary classification constraint and instead assigns a probability distribution of cluster centroids to features and observations to the dataset, hence the name "soft-clustering." Due to its flexibility, it has become a core part of the NLP toolkit, as it is designed to identify latent topics within a collection of discrete data, such as a corpus of text documents. In an NLP context, the LDA assumes that each document is a mixture of various topics and that each topic is characterized by a distribution over terms.

The standard LDA model has three parameters: the number of topics, K, the Dirichlet distribution prior on the per-document topic distribution, α and the Dirichlet distribution prior on the per-topic term distribution, η. While the LDA is an unsupervised algorithm, and as such the objective evaluation of model performance is less apparent than say in a regression model, there are some metrics that can guide selecting the right values of the parameters. Beyond hyperparameter tuning, theoretical considerations, subjective interpretability, as well as established literature can guide in selecting the right parameters. Following Hansen et al. (2018), I set $\alpha = 50/k$ and $\eta = 0.025$. The low η value promotes a sparse term distribution, as such topics are characterized by a limited set of prominent words, while values of $\alpha > 1$ promote a more uniform per-document topic distribution.

Two common methods of evaluating LDA-s are Perplexity and Coherence scores. Perplexity score measures how well the LDA model predicts the sample. Perplexity is calculated as $\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{d=1}^{D}\log P(v_d)\right)$, where N is the total number of words, D is the total number of documents and $P(v_d)$ is the probability of terms in document d. A lower Perplexity score is associated with better model performance. Coherence measures the the co-occurence of top words of each topic estimated by the model. Coherence[7] score is calculated as $\text{Coherence} = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\log\frac{D(v_i,v_j)+\epsilon}{D(v_i)}$, where $D(v_i, v_j)$ is the number of documents containing both terms v_i and v_j, and $D(v_i)$ is the number of documents

[6] The three cutoff values correspond to a 0.05%, 0.025% and 0.0125% relative document frequencies respectively.

[7] Coherence scores are calcualted on a per-topic basis, averaging them can suffice as an overall measure of coherence in the model.

containing term v_i. A higher Coherence score is associated with better interpretability.

In order to find the optimal number of topics, I initially estimated LDA models over a vector of $K = [2, 40]$ and calculated the Perplexity and Coherence scores. All models are estimated using Gibbs sampling with 100 iterations and a burnin of 10. Figure 2 plots these scores of models estimated with each K topics.

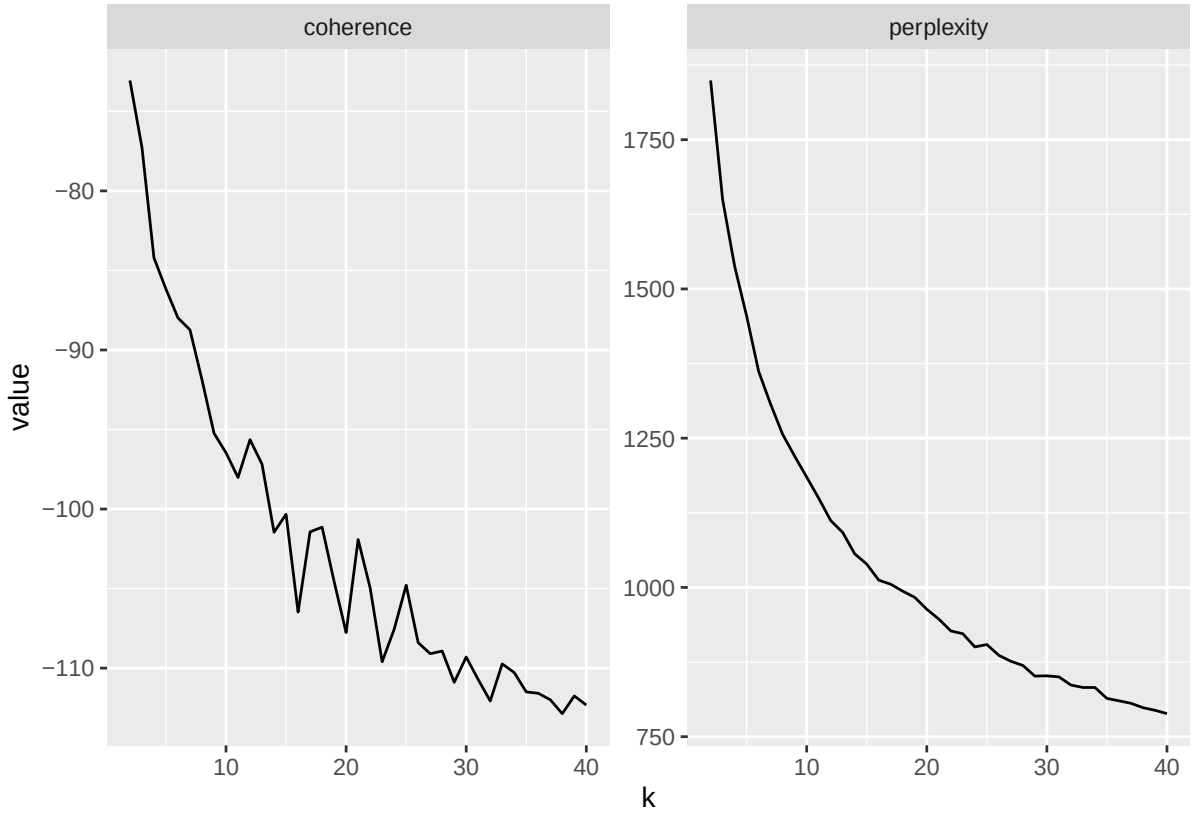

Figure 2: LDA evaluation metrics.

Unsurprisingly, aligned with the results in Hansen et al. (2018), Perplexity is constantly decreasing and promotes a larger number of topics. However, Coherence is also decreasing as the number of topics increases. This means that both promote the extremities on the possible values of K. Without a clear picture based on these scores, I default to using $K = 40$ as seen in Hansen et al. (2018).[8]

Using the LDA model, I estimate the posterior topic-term distributions. The interpretation of topics is subjective, and falls outside the scope of the LDA model, however, surprisingly, contradicting the results of the Coherence scores, the vast majority of topics can be easily interpreted. Figure 3 plots the top 12 words of each topic alongside with subjectively assigned thematic labels.

[8] Ez így nem jó, viszont a topic-ok javarészt könnyen értelmezhetőek. Ettől függetlenül mindenképp változtatok rajta, ezt és a következő fejezetet már csak lendületből írtam, módszerbeli átgondolás után megfeleltetve átírom majd.

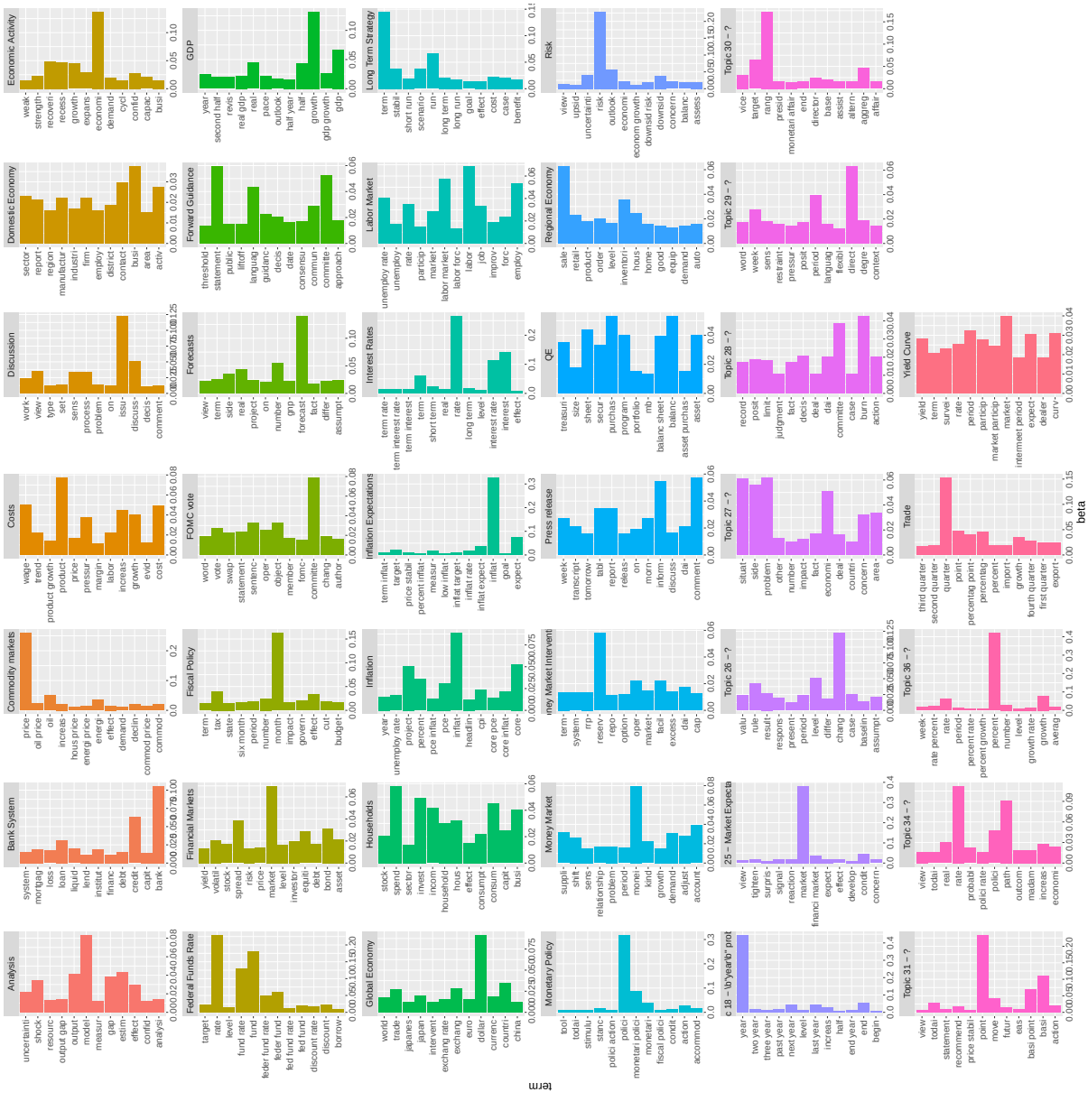

Figure 3: LDA evaluation metrics.

3 Empirical estimates

Coming soon...

4 Conclusions

Coming soon...

5 Egyéb konsziderációk (Jegyzet és ötlet outlet - to be removed)

- Valószínűleg "másolok" annyit a Hansen et al. (2018) cikk módszereiből, hogy a végső stádium környékén legalább egy e-mailt illene küldenem a szerzőknek.
- A témákat, szintén hasonlóan a Hansen et al. (2018) cikkhez érdemes lehet összevetni a Baker et al. (2016) indexeivel. Hansen-ék csak a main EPU-val vetik össze a téma distribution indexeiket, viszont itt egy kihagyott "ziccer" az EPU sub-indexekkel való összevetés. (Pl. Van EPU Monetary Policy, Fiscal Policy / Legislation, Financial Markets, stb index - a becsült témákkal ezek között könnyű párhuzamot vonni ezek között.)
- Sokkal több unique szavam van minden lépésben, mint a QJE-s Hansen et al. (2018) paper-nek. Ez nekem kétoldalról is alarming. Egyrészt, hogy nálam megy félre valami és ebből egy nagyon és ezért jönnek elő értelmezhetetlen témák. Ez kevésbé zavaró viszont, mivel a könnyen értelmezhető témák nagyon könnyedén érthetőek - pont mint náluk (sőt, még talán értelmezhetőbbek is). A másik, hogy nekik úgy van sokkal kevesebb, hogy valami ott nem jó a part-of-speech taggingben és igéik is maradnak, illetve náluk a raw szövegben is kevesebb a unique szó, mint ami nálam a szűrtben. Valamennyire érthető, ők egy subset-tel dolgoznak, viszont az a subset nagyságrendileg 70 százalékát lefedi a teljes transcripteknek - ehhez képest nálam a a szűrt adatban nálam kétszer annyi szó marad, mint náluk a nyersben. Vagy a FED írja le ennyire kevés szóval a közgazdaságtant, vagy valakinél valami félrecsúszott, ezt még nem tudom.
- Az LDA amit használok az egy unsupervised algoritmus, így mindig egyetlen feladata van, a megadott K klaszter-centroiddal megtalálni, hogy a belső heterogenitás mikor a legkisebb és az inter-klaszter heterogenitás mikor a legnagyobb - teljesen mindegy, hogy mik a centroidok (ezeket magának keresi meg). VISZONT, van ennek egy "semi-supervised" kiterjesztése, amiben előre lehet definiálni "guideline-okat" a klaszter centroidokhoz. Például: legyen az első téma az "inflation," és ekkor a topic 1-hez tartozó klaszter-centroid term-ek az "inflat," "cpi," "pce," "inflat rate," stb. Általában jó dolog, ha "szabadjára van engedve" és "let the data speak for itself," de ha nagyon nem akar összeállni a témák koherenciája, indokolt lehet a váltás erre. A másik érv e-mellett, hogy ez kivitelezhető közgazdasági elgondolás mentén. Pl.: A FED-et biztosan érdekli 1) az infláció, 2) a gazdasági teljesítmény, 3) a foglalkoztatás, 4) a pénzügyi rendszer stabilitása. Tehát kell lennie

legalább 4 klaszter központnak, amik néhány fogalommal viszonylag jól definiálhatók. Persze ez felveti a kérdést: akkor miért LDA és miért nem egyszerűen pre-definiált fogalmak mentén téma-proporciókat becsülni? Egyrészről, a többi szót ugyanúgy el kell helyezni és ezek aránya sokkal magasabb, mint a pre-definiáltaké. Másrészről, így megmarad az LDA soft-cluster előnye - tehát mind a pre-allokált, mind az algoritmus által allokált szavak "fluidak" maradnak a témák között. All-in-all pontosabban tudja lefogni, hogy a szövegekben miről mekkora arányban beszélnek. Hosszú okfejtés arról, hogy igazából mégsem értelmes, amit csináltam, de mindenesetre kíváncsi vagyok, hogy ezen érvek mellett szerinted értelmesebb-e így csinálni. (Ennek az átalakítása nem olyan "költséges," a kód pipeline adatfeldolgozó része borzalmasan hosszú, magát a modellt diagnosztikával és szubszekvent eloszlás becsléssel pár óra alatt újra lehet futtatni.)

- Offshoot ötlet: technikailag alkalmassá lehetne tenni a módszertant shock identifikációra. Aruoba and Drechsel (2024) egy újabb NBER-es publikáció ami text analysis alapú sokk identifikációt csinál, viszont ők csak a ColorBook-okat használják. A monetáris sokk "döntéshozási" szempontjából viszont fontos lehet a Transcriptekkel összevetni - vagyis Elemzői Vélekedés vs. Döntéshozói Vélekedés (sentiment) eltérése lehet egy kvázi sokk (vagy legalább is az identifikáció alapja).

- Offshoot2: Albrizio et al. (2023) vizsgál egy "attention to FED" hatást vállalati earningscall-okban. Ebből szép eredményeik vannak (segíti a transzmissziót, előrelátóbbak/kevésbé kitettek a sokkoknak, jobbak az inflációs várakozásaik). Erre esetleg lehetne építeni - szintén sentiment elven, FED Vélekedés vs. Szakmai News Article (jobb híjján Google News) Vélekedése.

6 References

Albrizio, Silvia, Allan Dizioli, and Pedro Vitale Simon. 2023. *Mining the Gap: Extracting Firms' Inflation Expectations from Earnings Calls*. International Monetary Fund.

Aruoba, S Borağan, and Thomas Drechsel. 2024. "Identifying Monetary Policy Shocks: A Natural Language Approach." National Bureau of Economic Research.

Baker, Scott R, Nicholas Bloom, and Steven J Davis. 2016. "Measuring Economic Policy Uncertainty." *The Quarterly Journal of Economics* 131 (4): 1593–1636.

Blei, David M, and John D Lafferty. 2009. "Topic Models." In *Text Mining*, 101–24. Chapman; Hall/CRC.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Caldara, Dario, and Matteo Iacoviello. 2022. "Measuring Geopolitical Risk." *American Economic Review* 112 (4): 1194–1225.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach." *The Quarterly Journal of Economics* 133 (2): 801–70.

Justeson, John S, and Slava M Katz. 1995. "Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text." *Natural Language Engineering* 1 (1): 9–27.

Romer, Christina D, and David H Romer. 2023. "Presidential Address: Does Monetary Policy Matter? The Narrative Approach After 35 Years." *American Economic Review* 113 (6): 1395–1423.

Wijffels, Jan. 2021. "Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the'UDPipe' 'NLP'toolkit. 2020." *R Package Version 0.8* 8.