CliqueCorex: A Self-supervised Clique-based Anchored Topic Model

Sami Diaf
Universität Hamburg
Department of Socioeconomics
sami.diaf@uni-hamburg.de

Abstract

Probabilistic generative topic models are the de facto choice for most text data applications, usually augmented with unsupervised and semi-supervised learning strategies to enhance the topic quality. Alternatively, information theory was used to build model-free algorithms able to learn homogeneous, binary latent groups of words, as topics, via multivariate mutual information as for the *Correlation Explanation* model (*CorEx*), with the possibility of incorporating anchors, or keywords, as prior information that better reflects the practitioner's experience to reveal nested topics. This paper establishes a self-supervised, anchor-based strategy, namely *CliqueCorex*, where anchors are meaningful subgraphs resulting from the hierarchical clustering of the corpus' bigrams via *clique percolation* algorithm. This scheme maximizes the information extraction by learning cohesive topics without biased prior information or any additional hyperparameter optimization. Applied to two central banking corpora, *CliqueCorex* improved the plain *CorEx* results without the need of additional topics, while uncovering nested topic contents, spanning across a wide spectrum of monetary policy practices, with a natural separability and an importance order that demonstrate the usefulness of cliques when implementing a guided inference.

1 Introduction

The abundance of textual sources and their growing complexity has led to continuous attempts to improve the existing text-as-data methods. These efforts have either sought sophistication from neural networks, or improved existing generative models to better handle the studied task (Churchill and Singh, 2022), at the expense of detailed hyperparameter specifications (Airoldi et al., 2014; Gallagher et al., 2017).

In machine learning, probabilistic topic models are still considered as the workhorse for most text mining applications, particularly the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) whose scheme has been adopted later by many strategies that brought improvements via adding metadata as covariates (Blei and McAuliffe, 2007), time-based topics (Blei and Lafferty, 2006) or nested hierarchies (Griffiths et al., 2003). Semi-supervised approaches (Lu et al., 2011; Jagarlamudi et al., 2012; Eshima et al., 2020) have been used in several applications, consisting of allowing practitioners to set prior lexical information, in the form of keywords or labels (Nomoto, 2022), as an attention mechanism to learn robust topics and test hypotheses in a guided fashion, although setting keywords is still considered as a rule-of-thumb exercise (King et al., 2017; Eshima et al., 2020).

While most bag-of-word topic models have been criticized for yielding poor results, due to count data whose structure ignores words' interactions, Steeg and Galstyan (2014) proposed to learn topics from a different perspective using information theory, by computing multivariate mutual information of relevant groups of words that form latent features, known as topics. Correlation Explanation (*CorEx*) (Steeg and Galstyan, 2014) has the advantage of being neither a generative model nor requiring assumptions, but still capable of uncovering meaningful features in diverse applications with sparse data. Extensions of CorEx offer a semi-supervised approach based on predefined keywords, or anchors, that translates the experience or beliefs of practitioners, to learn specific topics as well as hierarchical structures via chained inference (Gallagher et al., 2017).

In network analysis, practitioners usually aim to cluster data into homogeneous groups using several criteria, falling into the class of hierarchical clustering task. Blondel et al. (2008) and Traag et al. (2019) proposed optimized clustering schemes for nonoverlapping features known as *communities*, while Derényi et al. (2005) devised *clique per-*

colation to learn subgraphs as overlapping communities, known as *cliques*, later extended to unweighted and weighted graphs (Farkas et al., 2007). Examples of application in textual analysis considered words as nodes and used communities for document scaling (Diaf, 2023), rhetoric studies (Rule et al., 2015; Bail, 2016) and cliques for an efficient topic detection on short documents (Churchill and Singh, 2020).

If setting keywords improves greatly the quality of the inferred topics (Eshima et al., 2020), it remains mostly an unsupervised task (King et al., 2017), highly dependent on the experience and views of practitioners (Nomoto, 2022), which may show interests in small nested topics not often captured by the models. This work sets an automated anchor strategy, namely *CliqueCorex*, to select keywords, as meaningful features from clique percolation, and to feed them to an anchored CorEx (Gallagher et al., 2017) to maximize topic extraction. Thus, blending two separate clustering schemes into one self-supervised topic model capable of determining keyword groups without human intervention. The number of topics, considered as a hyperparameter, is automatically set to the number of cliques, although users can extend it to learn extra features not coined to any clique.

This strategy frees practitioners from analyzing the corpus in search of relevant keywords and instead automates it by uncovering meaningful mixed-membership cliques, whose semantic structure can be assimilated to powerful subtopics (Ohsawa et al., 1998) emerging from the corpus itself without requiring external prior information. I argue that cliques, as anchors, reinforce the learning process of *CorEx* models by detecting maximally informative latent groups of words, with a preserved semantic structure and an importance order. In other terms, setting semantically-grounded ngrams as prior acts as a semantic regularization to force a more context-based inference.

In many application fields, setting keywords remains a delicate task especially when a word is polysemic or linked to many topics. As for central banking communication, the word *rate* is used in three key measures of monetary policy practices (interest rate, inflation rate and unemployment rate) as well as in other technical terms. Assigning the word *rate* to a unique anchor group may be problematic in probabilistic topic models, as it could later appear in other topics not related the three aforementioned measures, hence lowering the topic quality of the learned models. *CliqueCorex* solves this issue by automatically setting anchors, as mixed-membership nodes, without any constraint on the cliques.

Applied to two different central banking corpora, *CliqueCorex* unfolded granular themes in the Federal Reserve (FED) governors' speeches (1996-2020), where topics revealed the importance of banking supervision and the macroeconomic status in the U.S. central banking discourse, along other secondary, but not less important interests for central bankers as for market competition and innovation. On the European Central Bank speeches (1997-2023), *CliqueCorex* revealed a policy-oriented corpus, closely tied to the ECB missions and objectives, with different interests and reduced topics compared to the FED corpus. In both applications, *CliqueCorex* outperformed standard *CorEx* in terms of total correlation and topic quality.

The paper outlines the build-up of *CliqueCorex* from a network analysis perspective and from statistical learning (Section 2), then implements the proposed algorithm on two central banking corpora (Section 3) and compares them to the standard *CorEx* used by practitioners.

2 Methodology

2.1 Clique Percolation

Uncovering homogeneous groups in a dense, heavily connected network is a difficult task that requires advanced techniques, exceeding the classic clustering methods as for Principal Component Analysis and the K-means.

In network analysis, we define a node as the representation of an entity or a word (Mihalcea and Tarau, 2004), while an edge connects two entities, either directed or indirected. If an edge determines the strength of the link between two nodes, then the network is said to be weighted, otherwise it is unweighted.

We refer to community detection (Fortunato, 2010) the process of identifying strongly connected subgraphs in a given network, usually assimilated to a hard clustering exercise, that assigns each node to just one community (Blondel et al., 2008; Traag et al., 2019), ignoring cases where a node could be shared by many communities, similar to soft clustering. This overlapping feature was devised as *clique percolation* for unweighted (Palla et al., 2005) and weighted (Farkas et al., 2007) graphs.

In text mining, community detection has been used in a nonoverlapping context to better scale documents (Diaf, 2023), extract meaningful word groups (Bail, 2016) or to study lexical shift (Rule et al., 2015), while attempts to use clique percolation targeted topic modeling for small documents (Churchill and Singh, 2020). Nonoverlapping structures are seen as independent features, but several words could belong to different communities, as for the word *united* shared by many entities (United States, United Kingdom, United Nations), hence the necessity to use clique percolation to take into account words' multiple memberships.

As given by Farkas et al. (2007), clique percolation first identifies k-cliques which are fully connected networks with k nodes (starting from k=3) and filters those having theirs intensities higher than a given threshold I. The intensity of a clique C, denoted I_C, is simply the geometric mean of the edge weights w_{ij} associated to the nodes i and j:

$$I_c = \left(\prod_{i<j;i,j \in C} w_{ij} \right)^{(2/k(k-1))}$$

An intermediate step was added to the clique percolation algorithm, known as *CFinder* (Adamcsek et al., 2006), consisting of applying the intensity threshold I to the overlapping cliques, in addition to the already existing k-cliques (Lange, 2021). The challenge here is to optimize both k and I so to not exclude too many nodes and not incorporate many of them. Because percolation assumes the size distribution of communities are following a power-law, the optimal I for each k is just the cut-off above the emergence of a gigantic component (Lange, 2021).

2.2 Correlation Explanation

Most of topic modeling algorithms belong to the generative class (Churchill and Singh, 2020), assuming that documents are generated by a known distribution of terms. Their inference optimizes parameters of topic/term distribution so to maximize likelihood of documents in the dataset over k topics. A popular example is the Latent Dirichlet Allocation (Blei et al., 2003) which sets the basis of most topic models built upon the bag-of-word assumption, with a probabilistic inference that ignores word associations.

Attempts to improve topics' quality led to explore information theory in seeking robust, highly informative groups of words that occur together, without the need to use probabilistic simulations. Steeg and Galstyan (2014) proposed to use *total correlation* (TC) as a measure of mutual information among many variables:

$$TC(X_G) = \sum_{i \in G} H(X_i) - H(X_G)$$

where $H(X) = E_X[-log_2 p(x)]$ is the entropy measure and G denotes a subset of X random variables, in our case words. The total correlation is non-negative and equals zero if and only if the probability distribution factorizes. It could be written as a Kullback-Leibler divergence:

$$TC(X_G) = D_{KL}(p(X_G) || \prod_{i \in G} p(x_i))$$

Searching for latent factor Y, with k possible values, that explains the correlation in X makes the optimization search over all probabilistic functions of X, $p(y|x)$, as :

$$\max_{p(y|x)} TC(X;Y) \text{ s.t. } |Y| = k$$

For m different factors of Y_i, the optimization of *CorEx* (Gallagher et al., 2017) is written as:

$$\max_{G_j, p(y_j|x_{G_j})} \sum_{j=1}^{m} TC(X_{G_j}; Y_j)$$

where Y_j are m binary latent features, or topics, having X_{G_j} as their corresponding groups of word types. Latent factors Y_j are optimized to be informative about dependencies in the data and do not require generative modeling assumptions (Gallagher et al., 2017). Once learned, they can be used iteratively to construct new latent factors in a hierarchical fashion.

The numerical optimization of *CorEx* begins with a randomly initialized parameters, later iteratively updated as for *Expectation-Maximization* algorithm (Gallagher et al., 2017), which adds a binary parameter $\alpha_{i,j}$ equaling one if and only if word X_i appears in topic Y_j (i.e. $i \in G_j$). The previous equation will have its constraint on non-overlapping groups transformed into α:

$$\max_{\alpha_{i,j}, p(y_j|x)} \sum_{j=1}^{m} \left(\sum_{i=1}^{n} \alpha_{i,j} I(X_i : Y_j) - I(X : Y_j) \right)$$
$$\text{s.t. } \alpha_{i,j} \mathbb{1}[j = arg\max_{\bar{j}} I(X_i : Y_{\bar{j}})]$$

where $\alpha \in [0,1]$ is updated at iteration t by $\alpha_{i,j}^t = exp(\lambda^t (I(X_i : Y_j) - \max_{\bar{j}}(X : Y_{\bar{j}})))$

with λ controlling the sharpness of the softmax function.

For *anchored CorEx*, the objective remains the same as for unsupervised *CorEx* but with the inclusion of Z as labels of X, so that the information bottleneck (Gallagher et al., 2017) could be written as:

$$\max_{p(y|x)} \beta I(Z : Y) - I(X : Y)$$

where β controls the trade-off between compressing X and preserving information about the relevance variable Z. Coining a single word X_i to a topic Y_j results in constraining the above optimization scheme by setting $\alpha_{i,j} = \beta_{i,j}$ where $\beta_{i,j} \geq 1$ controls the strength of the anchor. The scheme remains similar if many anchor words are assigned to a given topic (Gallagher et al., 2017).

Algorithm: CliqueCorex

1.Clique detection: Run *clique percolation* algorithm (Farkas et al., 2007) over the network of bigrams and extract k overlapping groups of words, or *cliques*, from bigrams whose occurrence in the corpus is greater than π.

2.Anchored CorEx: The k cliques are used as features to learn an anchored CorEx model (Gallagher et al., 2017) with k topics.

The proposed *CliqueCorex* blends an unsupervised clustering scheme on the corpus' bigrams to uncover cliques as cohesive subgraphs, then use them as robust anchors, or keywords, for the semi-supervised *CorEx*. This frees the application from human intervention and the need of a transfer learning scheme built on external source as for word embeddings (Mikolov et al., 2013; Dieng et al., 2019), specific indexing (Medelyan, 2009) or classifiers (Florescu and Jin, 2018). *CliqueCorex* has the advantage of being naturally tailored to the documents' specification by seeking most informative features in the corpus. Moreover, the number of anchors is expressed as the number of groups emerging from the clique percolation, where each group can have $k \geq 3$ terms, depending on the specifications used (Farkas et al., 2007).

This hybrid scheme transforms the semi-supervised anchored *CorEx* into a fully self-supervised topic model, where the identification of cliques helps relieving the bottleneck when compressing data X into a set of topics Y. Furthermore, the number of topics, equaling the number of retrieved cliques, could be reduced by running another *CorEx* pass, so to build a hierarchical *CorEx* (Gallagher et al., 2017), if the number of cliques is relatively high.

3 Application

3.1 FED speeches

1,488 governor speeches at the U.S. Federal Reserve, during the period 1996-2020, were scraped from the institution's website[1], offering historical developments that accompanied the American monetary authority throughout several episodes and crises over the last three decades. The corpus was lemmatized using *udpipe* model (Straka et al., 2016) to reduce the size of the document-term-matrix and to get robust ngrams when applying clique percolation. This yielded 36 cliques[2], mostly sequences of three words, to be given as keyword groups for the anchored CorEx[3].

Table 2 shows the clique percolation output, consisting of 36 cliques expressing core monetary and macroeconomic interests, along a technical jargon used to describe the economic status and market developments. Cliques contain mostly trigrams and are informative, in a sense that their structure is similar to subtopics. For context-rich terms, as for "*macroeconomic*" and "*inflation*", their associated cliques feature more than 3 terms.

Cliques of Table 2 are used to learn an anchored *CorEx* whose results are shown in Table 4. Topics are ranked by an descending order of importance, based on their contribution to the total correlation (TC), where the first two confirm the importance of banking supervision and macroeconomic stability when communicating about monetary policy in the United States. Efforts of stability and supervision are linked to the post-2008 addresses and interpreted as direct signals toward economic agents (EuropeanParliament et al., 2018).

The remaining topics are a mix of technical (topics 3, 5 and 10) and non-technical topics (topics 6, 7 and 13), the latter consist of a descriptive jargon used in standard economic and financial analyses. Furthermore, the last three topics (34, 35 and 36) having the least contribution to the total correlation are tied to the crisis time 2007-2009, dealing respectively with the housing market, securitization/debt and oil prices.

Noticeable is that *CliqueCorex* outperforms classic *CorEx* in terms of total correlation (Table 1) and topic content (Table 5) whose structure does not prioritize central banking jargon, but rather frequent terms appearing in the corpus. For instance, Topic 1 learned by *CorEx* refers to academic papers and other references used by governors, which similarly

[1] https://www.federalreserve.gov/newsevents/speeches.htm

[2] The application took into account cliques containing few words, for the sake of illustration.

[3] The learning process for both corpora used anchored *CorEx* with 100,000 iterations and an anchor strength β_{ij}=2

Table 1: Total Correlation (TC) yielded by each model for both corpora.

Corpus \ Model	Plain CorEx	CliqueCorex
FED speeches	162.63	177.61
ECB speeches	141.01	144.78

appears in the 17^{th} position in the *CliqueCorex* output (Topic 17).

3.2 ECB Speeches

Speeches of executive members at the ECB (2,493 addresses from February 1997 to September 2023) were collected from the ECB website[4] and underwent the same steps to extract cliques as for the FED corpus.

Table 3 shows clique percolation results, that yielded 26 cliques, mostly trigrams and not rich in information as for the FED corpus. Particularly, the ECB corpus does not feature a strong, colorful macroeconomic jargon, but instead stresses out country-related interests because of the monetary union context.

The application of the *CliqueCorex* has a better total correlation than a plain *CorEx* (Table 1) as well as a better topic content, which turned out to be highly policy-oriented and lacking macroeconomic aspects as found in the FED corpus. This is most likely due to the broad interests discussed at the monetary union level, not at a country-specific level as for the FED. For instance, fiscal concerns linked to crises experienced by some country members (Topics 7 and 8 in Table 6) were broadly debated then economic indicators and forecasted aggregates (Topic 12).

Key terms like "*macroeconomic*" and "*inflation*" are not uncovered in the clique percolation step and only "*inflation*" is captured later by *CliqueCorex* in a more structural context involving unemployment and growth (Topic 21 in Table 6), but far less important then other descriptive topics.

Moreover, the corpus contains small multilingual paragraphs used by some speakers during their addresses (Topic 17 in Table 6 and Topic 2 in Table 7) and specific interests other than its main mission of price stability.

4 Conclusion

Probabilistic topic models continue to be the go-to solution when dealing with textual data under its different aspects. Behind their popularity, they are still limited by the fact that they consider words as independent features, in addition of having a probabilistic learning process based on word counts, yielding poorly informative results. While numerous extensions were developed to improve topic extraction, developments based on information theory suggest model-free algorithms capable of yielding superior topic quality, as for *correlation explanation*, and offering semi-supervised extensions for keywords and labels. This paper proposed the use of *clique percolation*, as a pre-processing step, enabling the automatic identification of anchors as *cliques* for a fully-automated anchored *CorEx*, encompassing neither external information nor human intervention. The resulting blend, named *CliqueCorex*, is a self-supervised topic model built on subgraphs, assumed to be subtopics from the corpus' bigram network, yielding maximally informative binary topics when blended to an anchored *CorEx*.

By adopting cliques as a semantic regularization scheme, *CliqueCorex* proved a higher ability in capturing hidden topics and other features overlooked by non-guided topic models. On two central banking corpora, known to have a rich imbricated context, *CliqueCorex* demonstrated a robustness in unfolding deep interests in monetary policy practices and reveal their relative importance, with rich monetary policy-oriented topics found at the U.S. Federal Reserve addresses, while policy-oriented interests dominate the speeches given at the European Central Bank, but not necessarily monetary or macroeconomic ones.

Clique percolation first revealed different scopes of interests the corpora have, although both deal with monetary policy, and the necessity to set tailored anchors for each corpus. This reinforces the claim that transferring keywords or anchors within the same task is not always indicated to extract nested features.

Uncovered cliques, in addition of acting as subtopics because of their semantic structures, fitted perfectly the anchored *CorEx* mechanism to deliver cohesive topic contents and reveal interesting corpus' orientation, in terms of topic content and importance.

[4] https://www.ecb.europa.eu/press/key/html/downloads.en.html

Table 2: Cliques in the FED corpus

Clique	Words
1	bank thrift supervision
2	macroeconomic policy stability objective
3	economic growth prospect
4	free trade flow
5	long time horizon
6	less well able
7	let now turn
8	dual mandate objective
9	climate risk relate
10	term treasury yield
11	across many market
12	achieve domestic inflation run trend goal objective
13	one part time
14	one recent study year
15	debt service burden
16	across national border
17	discussion paper series
18	put upward pressure
19	early first half last next past several ten three time twenty two week year
20	core pce price
21	committee fomc member participant
22	another important key reason way area
23	real short term time
24	government guarantee program
25	federal government spending
26	consumer financial protection
27	reduce regulatory burden
28	find new way
29	asset liability side
30	even far great
31	base capital measure
32	american community economic
33	develop new world
34	consumer durable good
35	agency debt issue security mbs
36	high oil price

Table 3: Cliques in the ECB corpus

Clique	Words
1	full information set
2	just mention two
3	precise quantitative definition
4	policy relevant horizon
5	billion euro banknote coin
6	content presentation slide
7	automatic fiscal stabiliser
8	greece ireland portugal
9	france germany italy
10	credit loss provision
11	conference discussion paper series
12	commission economic forecast
13	area non resident
14	available empirical evidence
15	already know well
16	common different objective set
17	council decide last meet
18	analysis can find
19	asset portfolio allocation
20	annual data report requirement
21	high structural unemployment
22	council decision make take
23	general term orientation
24	single technical platform
25	become self evident
26	become fully operational

Table 4: *CliqueCorex* on the Federal Reserve speeches

Topic	Top Words
1	supervision, regulator, supervisor, supervisory, banking, regulation, institution, deposit, oversight, bank
2	stability, macroeconomic, policy, shock, central, international, policymaker, objective, crisis, implication
3	productivity, growth, output, labor, production, gdp, prospect, worker, boost, wage
4	trade, flow, free, exchange, foreign, dollar, denominate, asia, currency, asian
5	horizon, argue, long, theory, empirical, time, weight, variable, argument, optimal
6	less, able, well, same, without, amount, net, account, only, both
7	turn, now, let, then, think, second, give, out, after, begin
8	mandate, dual, objective, laubach, bind, woodford, deviation, curve, reifschneider, jackson
9	risk, relate, certain, must, subject, procedure, whether, function, allow, exercise
10	treasury, yield, term, fund, return, normal, bond, condition, maturity, back
11	across, many, result, example, market, process, exist, require, significant, limit
12	inflation, run, outlook, monetary, employment, unemployment, nominal, trend, consumption, forecast
13	time, part, one, if, under, could, some, activity, case, because
14	recent, factor, percent, study, datum, one, compare, indicate, year, survey
15	service, industry, technology, customer, competitive, competition, innovation, technological, marketplace, efficiency
16	border, across, national, cross, country, globalization, among, infrastructure, nation, western
17	paper, pp, series, vol, journal, economics, pdf, discussion, cambridge, washington
18	pressure, upward, put, downward, demand, ease, japan, recovery, cut, stimulus
19	early, three, two, past, first, half, several, time, last, year
20	price, pce, core, estimate, indicator, historical, food, gradual, projection, solid
21	fomc, committee, open, accommodation, target, maximum, stance, incoming, path, easing
22	key, approach, area, discuss, reason, framework, analysis, practice, specific, implement
23	short, term, real, suggest, appear, relative, period, likely, somewhat, time
24	government, guarantee, program, reform, taxpayer, brothers, bankruptcy, street, suffer, serious
25	spending, expenditure, fall, household, percentage, down, index, quarter, recession, sustainable
26	loan, lending, borrower, credit, protection, lender, access, mortgage, consumer, commercial
27	regulatory, requirement, propose, organization, disclosure, proposal, compliance, examination, profile, rulemaking
28	find, way, question, problem, new, form, try, often, go, know
29	asset, liability, portfolio, investor, sheet, liquid, fail, instrument, against, arise
30	far, even, seem, indeed, great, moreover, little, still, quite, hence
31	measure, capital, thus, however, potential, base, reflect, generally, substantial, expect
32	community, american, education, family, school, training, americans, million, educational, census
33	world, develop, century, new, modern, history, society, power, dramatic, yet
34	consumer, home, homeowner, housing, income, residential, construction, foreclosure, homeownership, affordable
35	liquidity, agency, debt, loss, security, funding, counterpartie, stress, default, securitization
36	price, oil, decline, rise, sharp, high, supply, above, low, tighten

Table 5: Plain *CorEx* on the Federal Reserve speeches

Topic	Top Words
1	pp, vol, economics, journal, cambridge, david, massachusetts, nber, papers, university
2	spending, output, decline, rise, slow, boost, gdp, consumption, fall, labor
3	supervisor, supervisory, regulator, supervision, regulatory, regulation, oversight, exposure, banking, institution
4	fomc, inflation, outlook, monetary, employment, unemployment, expectation, nominal, stance, maximum
5	liquidity, systemically, funding, stress, severe, crisis, macroprudential, collateral, vulnerability, repurchase
6	text, pdf, speech, washington, governors, april, www, november, march, february
7	outsourcing, amy, needs, send, thankfully, retraining, multinational, involuntarily, saez, ashenfelter
8	organization, establish, transaction, develop, create, legal, act, protection, address, effort
9	expect, balance, risk, measure, potential, firm, asset, activity, sheet, generally
10	reduce, effect, relatively, short, investor, far, lead, capital, substantial, increase
11	practice, certain, apply, problem, procedure, limit, rule, issue, subject, set
12	macroeconomic, theory, policymaker, empirical, macroeconomics, taylor, equilibrium, influence, economist, variable
13	international, united, european, states, japan, foreign, exchange, central, global, currency
14	raghuram, rajan, gaps, makeup, text5, text6, text8, text7, text9, firms
15	education, educational, school, young, college, life, skill, americans, adult, population
16	production, productivity, equipment, worker, half, living, producer, inventory, war, fast
17	loan, borrower, lending, lender, mortgage, credit, commercial, underwriting, subprime, securitization
18	keys, toggle, caption, mediainfo, xs, transcripttext, transcriptlinkurl, col, fullscreen, myplayer
19	approach, framework, implement, assessment, appropriate, guidance, quantitative, conduct, consider, model
20	question, think, answer, reason, fact, argument, political, try, hand, precisely
21	housing, income, family, home, household, homeowner, residential, construction, homeownership, survey
22	shock, uncertainty, run, imply, term, normal, episode, movement, uncertain, response
23	suggest, period, argue, early, evidence, view, course, factor, adjust, quite
24	century, old, free, society, generation, nineteenth, history, virtually, twentieth, revolution
25	process, individual, information, involve, enhance, effective, ability, use, recognize, responsibility
26	technology, computer, technological, electronic, internet, service, network, automate, telecommunication, physical
27	actor, head, boivin, bridgewater, scrapping, scandinavian, risks, uncorrelated, confidently, vestin
28	debt, investment, finance, bond, borrowing, private, borrow, dollar, collapse, saving
29	audit, auditor, privacy, laundering, payments, sarbanes, oxley, sponsoring, treadway, merchant
30	analysis, example, datum, include, base, determine, particularly, study, relationship, relate
31	census, urban, local, racial, metropolitan, finances, barrier, resident, hispanic, hispanics
32	clive, bring, timmermann, penalver, billi, exploration, farmer, matheson, sandri, cardia
33	competitive, industry, competition, innovation, law, marketplace, efficient, opportunity, efficiency, competitor
34	hear, fed, listen, chair, district, communications, hope, proud, president, prepare
35	error, rational, rigidity, rose, agent, al, manuscript, et, override, linear
36	increasingly, facilitate, pricing, expand, availability, cash, pay, profitable, segment, hedge

Table 6: *CliqueCorex* on the ECB speeches

Topic	Top Words
1	information, set, introduction, regard, order, respect, importance, practice, specific, conduct
2	mention, two, seem, try, difficult, maker, experience, agent, another, world
3	definition, quantitative, precise, variable, reference, reserve, strategy, signal, influence, interpret
4	horizon, response, policy, underlie, relevant, volatility, imply, recent, associate, broad
5	banknote, coin, cash, changeover, circulation, card, store, january, dollar, electronic
6	slide, content, presentation, annex, kb, pdf, peter, flatten, download, proxy
7	fiscal, deficit, stabiliser, reform, pact, automatic, budget, sustainable, government, competitiveness
8	ireland, portugal, greece, spain, education, mobility, young, cyprus, belgium, slovenia
9	germany, italy, france, german, di, age, inequality, discourage, italian, five
10	credit, loss, systemic, crisis, banks, supervisory, supervision, liquidity, exposure, provision
11	paper, journal, vol, pp, economics, series, nber, university, research, al
12	forecast, economic, average, period, factor, gdp, estimate, percentage, indicator, hicp
13	non, investment, large, capital, reduce, small, united, turn, less, total
14	empirical, evidence, suggest, relative, effect, aggregate, shock, likely, premium, cycle
15	know, think, often, even, little, say, go, already, question, well
16	different, set, objective, common, principle, problem, framework, rule, define, task
17	decide, council, die, le, ich, der, la, zu, und, meet
18	analysis, find, theory, federal, model, argue, academic, behaviour, economist, bubble
19	asset, lend, portfolio, loan, sheet, bond, fund, maturity, household, yield
20	data, report, type, source, requirement, publish, distribution, disclosure, annual, sample
21	unemployment, inflation, structural, outlook, wage, decline, pressure, growth, low, rate
22	treaty, decision, independence, maastricht, council, union, state, mak, responsibility, institutional
23	general, term, short, degree, however, development, rather, similar, extent, although
24	border, single, platform, technical, infrastructure, field, service, efficient, payment, integration
25	become, self, cause, back, face, put, consequence, prevent, happen, bad
26	become, fact, example, role, base, reason, limit, fully, consider, conclusion

Table 7: Plain *CorEx* on the ECB speeches

Topic	Top Words
1	journal, paper, pp, economics, vol, nber, university, american, literature, al
2	die, der, und, ich, le, zu, je, la, auf, une
3	even, different, fact, seem, less, rather, little, case, reason, often
4	inflation, outlook, pressure, medium, wage, expectation, decline, projection, upward, uncertainty
5	relative, suggest, evidence, effect, factor, empirical, aggregate, likely, short, tend
6	rate, growth, structural, low, grow, gdp, inter, fiscal, condition, price
7	political, always, go, simply, idea, world, true, bad, must, perhaps
8	recovery, purchase, accommodative, household, ease, return, negative, stimulus, accommodation, fall
9	sheet, lend, fund, banks, loan, liquidity, crisis, credit, maturity, sovereign
10	leverage, asset, loss, response, episode, risk, macro, mitigate, boom, buffer
11	infrastructure, payment, transaction, user, settlement, provider, service, retail, field, initiative
12	pandemic, covid, lagarde, coronavirus, christine, digital, op, pepp, carbon, green
13	academic, theoretical, understand, press, economist, hypothesis, assumption, attempt, keynesian, conference
14	supervision, supervisory, systemic, regulatory, supervisor, basel, prudential, regulation, exposure, management
15	introduction, development, regard, general, thus, final, several, various, particular, mention
16	series, estimate, percentage, chart, income, bulletin, zero, survey, occasional, total
17	labour, productivity, export, labou, population, worker, capita, deficit, competitiveness, education
18	currency, banknote, changeover, exchange, coin, quot, accession, shall, enlargement, erm
19	value, refer, available, amount, type, use, participant, feature, form, wide
20	framework, authority, institution, principle, rule, task, procedure, implementation, set, establish
21	definition, analysis, reference, information, variable, conduct, quantitative, assessment, strategy, appropriate
22	integration, border, competition, cross, single, integrate, efficient, barrier, europe, transfer
23	treaty, pact, independence, emu, maastricht, credibility, institutional, union, decision, responsibility
24	fan, gov, villier, eride, ottaviano, wogau, hicpx, smile, disorder, spot
25	technology, happen, get, big, revolution, pay, online, protect, stop, era
26	gonzález, páramo, manuel, josé, pedersen, reluctant, responses, sifi, workshop, metrick

References

Balázs Adamcsek, Gergely Palla, Illés J. Farkas, Imre Derényi, and Tamás Vicsek. 2006. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021–1023.

Edoardo M. Airoldi, David Blei, Elena A. Erosheva, and Stephen E. Fienberg, editors. 2014. *Handbook of Mixed Membership Models and Their Applications*. Chapman & Hall / CRC Handbooks of Modern Statistical Methods. Taylor and Francis, Hoboken.

Christopher A. Bail. 2016. Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences*, 113(42):11823–11828.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 113–120, New York, NY, USA. Association for Computing Machinery.

David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 121–128, Red Hook, NY, USA. Curran Associates Inc.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Rob Churchill and Lisa Singh. 2020. Percolation-based topic modeling for tweets. *WISDOM 2020 : The 9th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining*.

Rob Churchill and Lisa Singh. 2022. The evolution of topic modeling. *ACM Comput. Surv.*, 54(10s).

Imre Derényi, Gergely Palla, and Tamás Vicsek. 2005. Clique percolation in random networks. *Phys. Rev. Lett.*, 94:160202.

Sami Diaf. 2023. CommunityFish: A Poisson-based document scaling with hierarchical clustering. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 59–67, Online. Association for Computational Linguistics.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019. Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907*.

Shusei Eshima, Kosuke Imai, and Tomoya Sasaki. 2020. Keyword assisted topic models. *arXiv preprint arXiv:2004.05964*.

EuropeanParliament, Directorate-General for Internal Policies of the Union, P Hubert, and C Blot. 2018. *Central bank communication during normal and crisis times – Monetary dialogue September 2018 – In-depth analysis*. European Parliament.

Illés Farkas, Dániel Ábel, Gergely Palla, and Tamás Vicsek. 2007. Weighted network modules. *New Journal of Physics*, 9(6):180.

Corina Florescu and Wei Jin. 2018. Learning feature representations for keyphrase extraction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Santo Fortunato. 2010. Community detection in graphs. *Physics Reports*, 486(3):75–174.

Ryan J. Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2017. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542.

Alan Greenspan. 2004. Risk and uncertainty in monetary policy. *American Economic Review*, 94(2):33–40.

Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. 2003. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213, Avignon, France. Association for Computational Linguistics.

Gary King, Patrick Lam, and Margaret Roberts. 2017. Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, 61(4):971–988.

Jens Lange. 2021. Cliquepercolation: An r package for conducting and visualizing results of the clique percolation network community detection algorithm. *Journal of Open Source Software*, 6:3210.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin K. Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 81–88.

Olena Medelyan. 2009. *Human-competitive automatic topic indexing*. Ph.D. thesis, Citeseer.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tadashi Nomoto. 2022. Keyword extraction: a modern perspective. *SN Computer Science*, 4(1):92.

Yukio Ohsawa, Nels E. Benson, and Masahiko Yachida. 1998. Keygraph: automatic indexing by co-occurrence graph based on building construction metaphor. *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries -ADL'98-*, pages 12–18.

Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

Alix Rule, Jean-Philippe Cointet, and Peter S. Bearman. 2015. Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, 112(35):10837–10844.

Greg Ver Steeg and Aram Galstyan. 2014. Discovering structure in high-dimensional data through correlation explanation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 577–585, Cambridge, MA, USA. MIT Press.

Milan Straka, Jan Hajič, and Jana Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1).