Text Mining: Basic Models and Applications

Alvaro J. Riascos Villegas
University of los Andes and Quantil

July 6 2018

Contents

Introduction

- Sixty years of efforts to model natural language have started to give useful results: Complete industries have build on top of the field of natural language processing.
- These models are meant to make natural language understandable to computers.
- For the first time we have methods to process automatically large amounts of text and extract different features that reveal syntactic and semantic characteristic of communication.

Introduction

- Sixty years of efforts to model natural language have started to give useful results: Complete industries have build on top of the field of natural language processing.
- These models are meant to make natural language understandable to computers.
- For the first time we have methods to process automatically large amounts of text and extract different features that reveal syntactic and semantic characteristic of communication.

Introduction

- Sixty years of efforts to model natural language have started to give useful results: Complete industries have build on top of the field of natural language processing.
- These models are meant to make natural language understandable to computers.
- For the first time we have methods to process automatically large amounts of text and extract different features that reveal syntactic and semantic characteristic of communication.

Linguistic model applications

- Vectorization of words or documents (general purpose).
- Sentiment analysis.
- Topic models.
- Automatic summaries.
- Document distance.
- Error correction models.

Contents

Vectorization

- Bag of words (represented as a Term Document Matrix)

doc1	Two for tea and tea for two
doc2	Tea for me and tea for you
doc3	You for me and me for you

	two	tea	me	you
doc1	2	2	0	0
doc2	0	2	1	1
doc3	0	0	2	2

- The Inverse Document Frequency Matrix weights each term t by $log\left(\frac{D}{D_t}\right)$ where D is the total number of documents and D_t the total number of documents in which t occurs.

Basics

- A Vocabulary V is a finite set of words.
- A Document $\mathbf{w}$ over V is $\mathbf{w} = (w_1, \ldots, w_n)$ with $w_i \in V$.
- A Corpus is a set of documents: $D = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$

Contents

Latent Dirichlet Allocation (LDA)

- David M. Blei, Andrew Y. Ng, Michael I. Jordan. 2003. Journal of Machine Learning Research 3, 993-1022.
- LDA is a generative probabilistic model of a corpus.
- The idea is that documents are represented by random mixtures of latent topics where each topic is a distribution of terms.
- It is a unsupervised technique that can be used for: topic modelling, feature extraction, collaborative filtering, modelling of categorical variables, etc.

Latent Dirichlet Allocation (LDA)

- David M. Blei, Andrew Y. Ng, Michael I. Jordan. 2003. Journal of Machine Learning Research 3, 993-1022.
- LDA is a generative probabilistic model of a corpus.
- The idea is that documents are represented by random mixtures of latent topics where each topic is a distribution of terms.
- It is a unsupervised technique that can be used for: topic modelling, feature extraction, collaborative filtering, modelling of categorical variables, etc.

Latent Dirichlet Allocation (LDA)

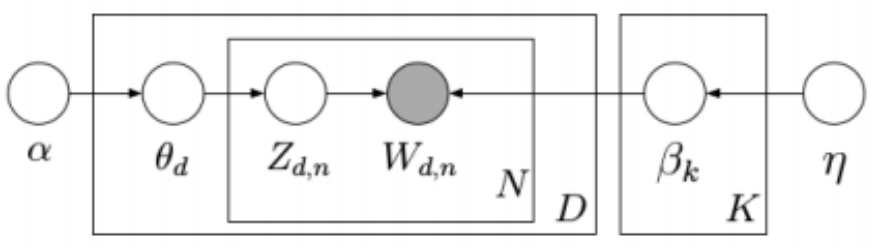
- David M. Blei, Andrew Y. Ng, Michael I. Jordan. 2003. Journal of Machine Learning Research 3, 993-1022.
- LDA is a generative probabilistic model of a corpus.
- The idea is that documents are represented by random mixtures of latent topics where each topic is a distribution of terms.
- It is a unsupervised technique that can be used for: topic modelling, feature extraction, collaborative filtering, modelling of categorical variables, etc.

Latent Dirichlet Allocation (LDA)

- David M. Blei, Andrew Y. Ng, Michael I. Jordan. 2003. Journal of Machine Learning Research 3, 993-1022.
- LDA is a generative probabilistic model of a corpus.
- The idea is that documents are represented by random mixtures of latent topics where each topic is a distribution of terms.
- It is a unsupervised technique that can be used for: topic modelling, feature extraction, collaborative filtering, modelling of categorical variables, etc.

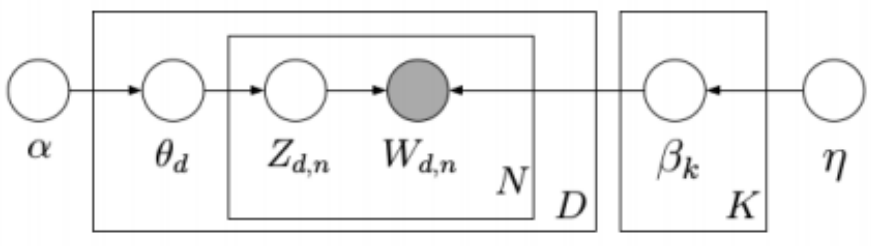
The probabilistic model

- Let **w** be a document of corpus D:
  1. Choose $N \sim \text{Poisson}(\xi)$.
  2. Choose $\theta \sim \text{Dirichlet}(\alpha)$.
  3. For each one of the N terms w_n:
     - Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
     - Choose a term w_n de $p(w_n \mid z_n, \beta)$, a multinomial distributions conditional on topic z_n, where β is a matrix of $k \times V$ and $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$.
- The objective is to estimate the distribution of words conditional on topics (β).



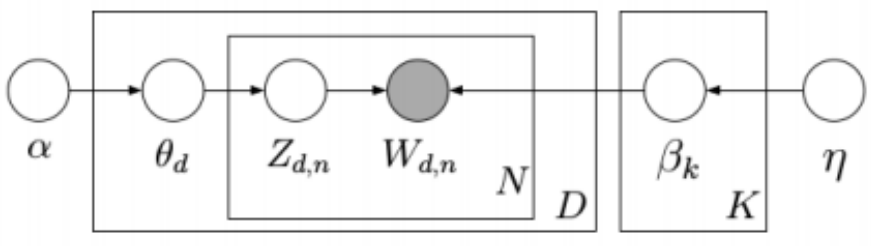
The probabilistic model

- Let **w** be a document of corpus D:
  1. Choose $N \sim \text{Poisson}(\xi)$.
  2. Choose $\theta \sim \text{Dirichlet}(\alpha)$.
  3. For each one of the N terms w_n:
     - Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
     - Choose a term w_n de $p(w_n \mid z_n, \beta)$, a multinomial distributions conditional on topic z_n, where β is a matrix of $k \times V$ and $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$.
- The objective is to estimate the distribution of words conditional on topics (β).



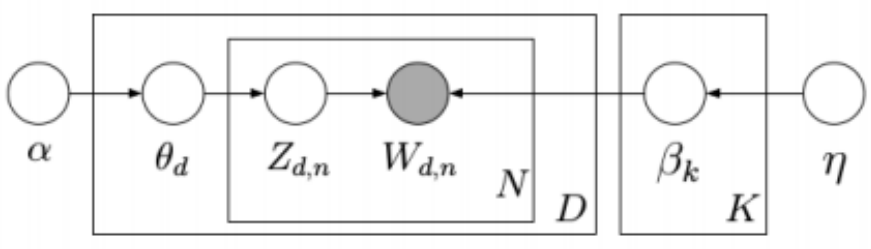
The probabilistic model

- Let **w** be a document of corpus D:
  1. Choose $N \sim \text{Poisson}(\xi)$.
  2. Choose $\theta \sim \text{Dirichlet}(\alpha)$.
  3. For each one of the N terms w_n:
     - Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
     - Choose a term w_n de $p(w_n \mid z_n, \beta)$, a multinomial distributions conditional on topic z_n, where β is a matrix of $k \times V$ and $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$.
- The objective is to estimate the distribution of words conditional on topics (β).



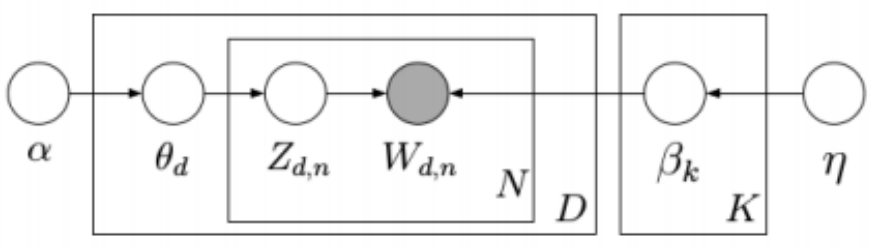
The probabilistic model

- Let **w** be a document of corpus D:
  1. Choose $N \sim \text{Poisson}(\xi)$.
  2. Choose $\theta \sim \text{Dirichlet}(\alpha)$.
  3. For each one of the N terms w_n:
     - Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
     - Choose a term w_n de $p(w_n \mid z_n, \beta)$, a multinomial distributions conditional on topic z_n, where β is a matrix of $k \times V$ and $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$.
- The objective is to estimate the distribution of words conditional on topics (β).



The probabilistic model

- Let **w** be a document of corpus D:
  1. Choose $N \sim \text{Poisson}(\xi)$.
  2. Choose $\theta \sim \text{Dirichlet}(\alpha)$.
  3. For each one of the N terms w_n:
     - Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
     - Choose a term w_n de $p(w_n \mid z_n, \beta)$, a multinomial distributions conditional on topic z_n, where β is a matrix of $k \times V$ and $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$.
- The objective is to estimate the distribution of words conditional on topics (β).



Estimation

- It is possible to set up the likelihood function of the corpus and after a lot of hard computational work, obtain the maximum likelihood estimation of all parameters.
- In particular we can estimate the distribution of topics.

Estimation

- It is possible to set up the likelihood function of the corpus and after a lot of hard computational work, obtain the maximum likelihood estimation of all parameters.
- In particular we can estimate the distribution of topics.

Topic Modelling: Gapmaps

http://seixas.quantil.co/gapmaps/api/site/

Topic Modelling

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Validation: Perplexity

- A common measure to evaluate and compare models is the perplexity:

$$Perplexity(D) = \exp(-\frac{\sum_{d=1}^{M} log(p(\mathbf{w}_d))}{\sum_{d=1}^{M} N_d})$$

Document Modelling

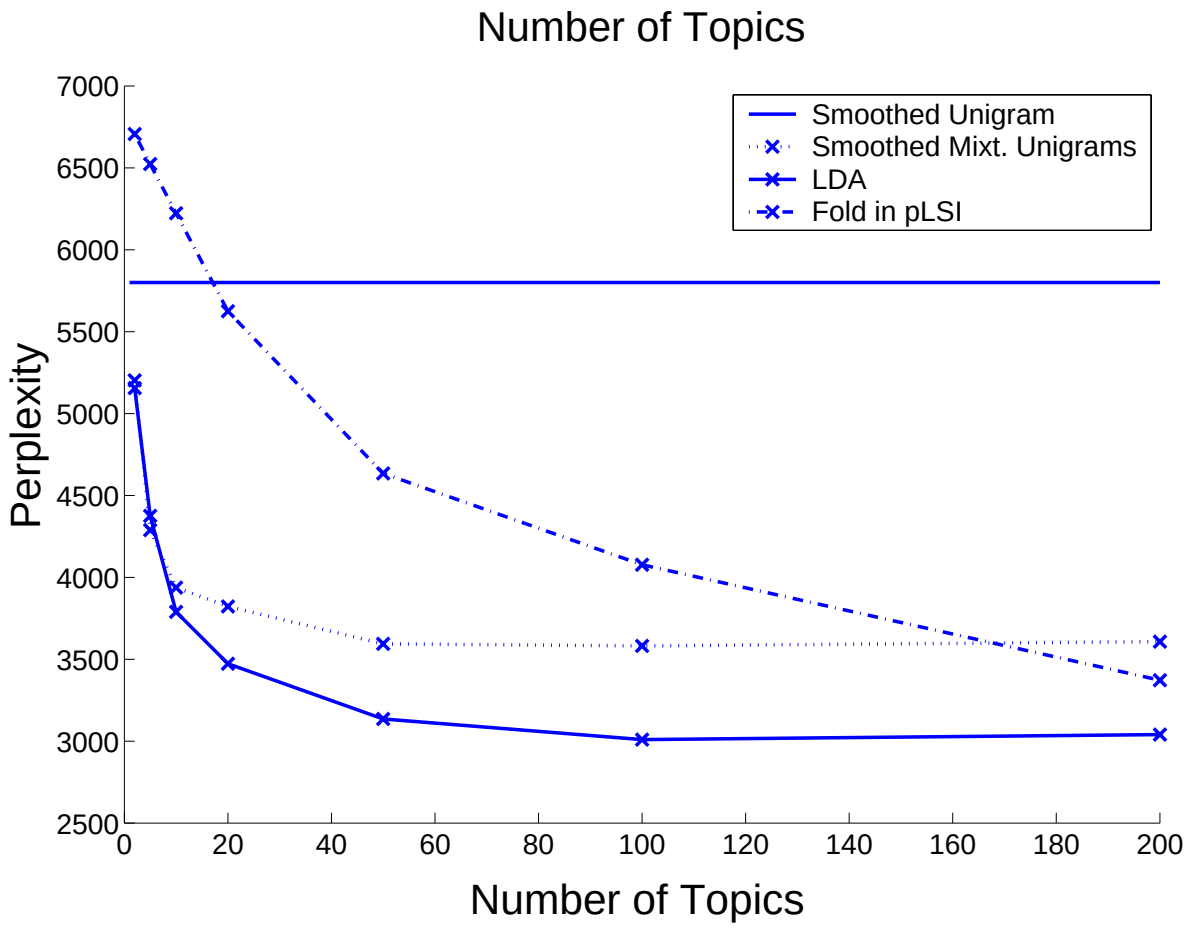


Figure 9: Perplexity results on the nematode (Top) and AP (Bottom) corpora for LDA, the unigram model, mixture of unigrams, and pLSI.

Classification using LDA features

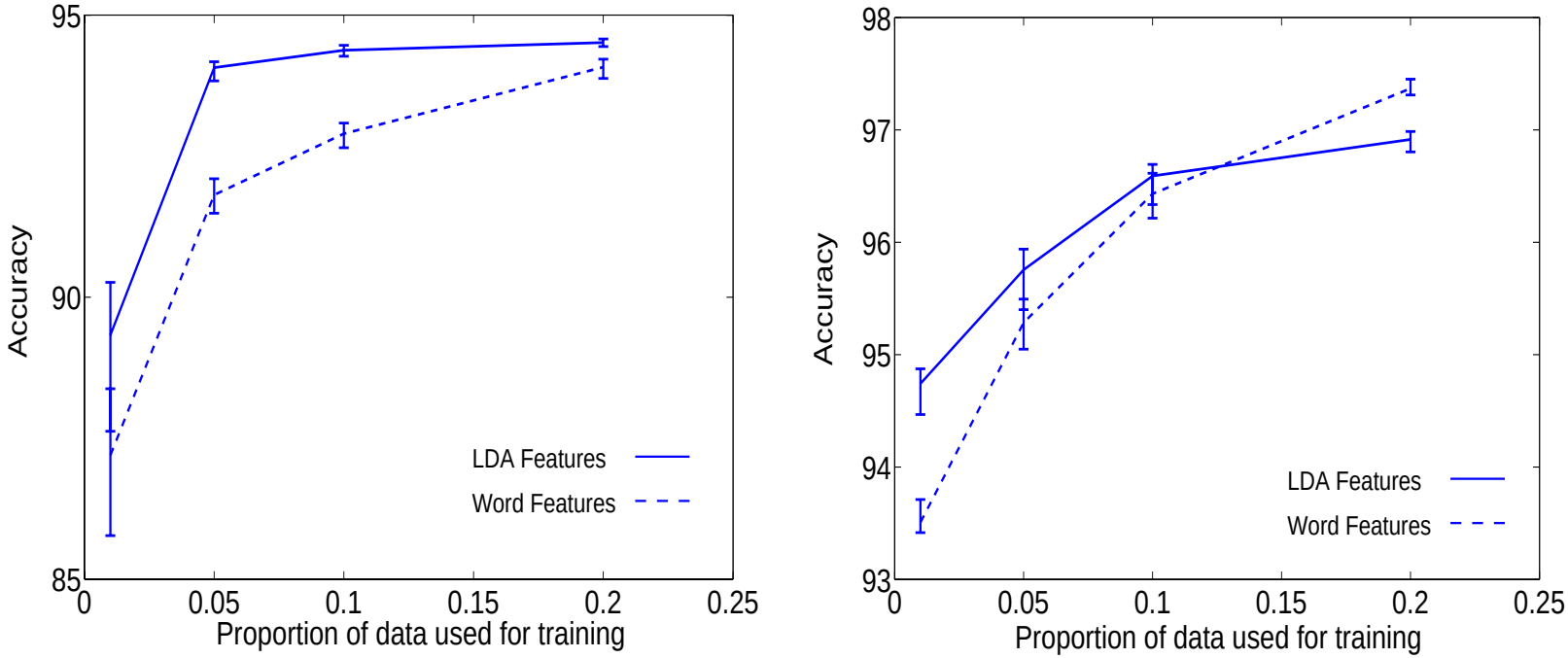


Figure 10: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN vs. NOT EARN. Graph (b) is GRAIN vs. NOT GRAIN.

Collaborative filtering

- In this case a set of users ranks a collection of films. Users are documents and films are words.

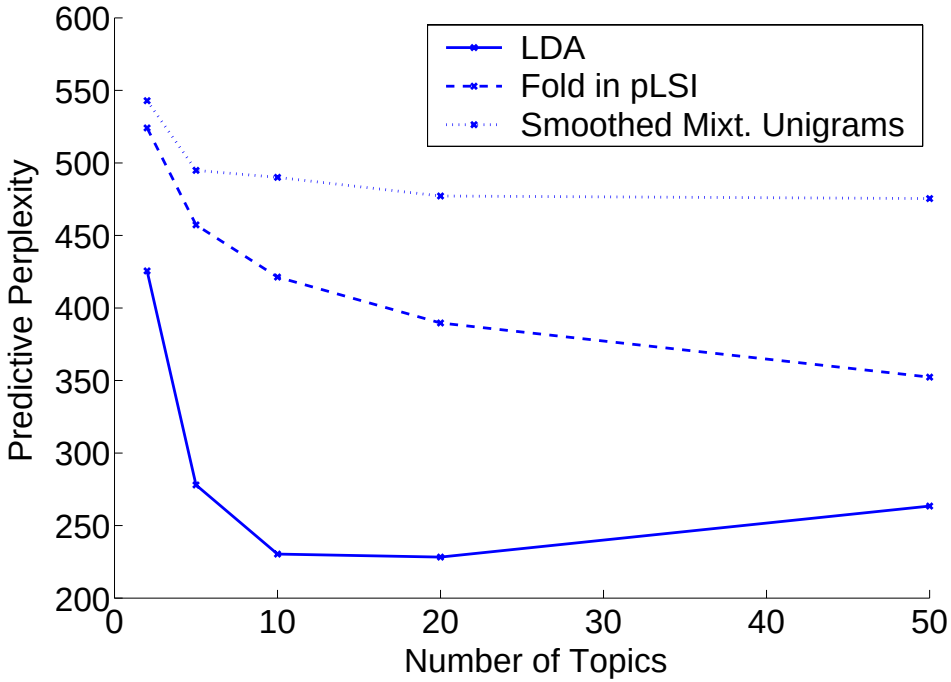
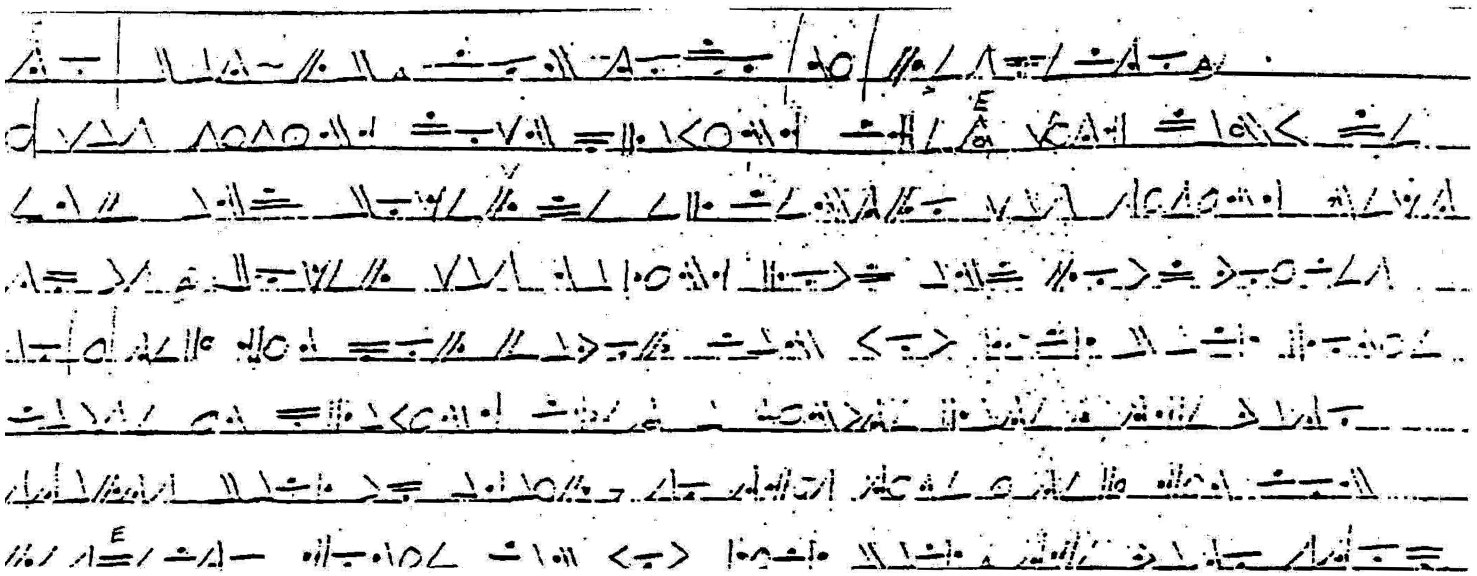


Figure 11: Results for collaborative filtering on the EachMovie data.

Contents

The Prisons Problem



Setting up the problem

- Students formalized the problem as one in which we need to find:

$$f : \{\text{Message symbols}\} \rightarrow \{\text{Traditional alphabet}\}$$

- The number of possibilities for f is enormous.
- To judge how plausible was a particular f they introduce a linguistic model.
- They took War and Peace and estimated the conditional frequency of letters.
- $M(x, y)$ is the conditional frequency of observing the word y given that you observed x immediately before.

Setting up the problem

- Students formalized the problem as one in which we need to find:

$$f : \{\text{Message symbols}\} \to \{\text{Traditional alphabet}\}$$

- The number of possibilities for f is enormous.
- To judge how plausible was a particular f they introduce a linguistic model.
- They took War and Peace and estimated the conditional frequency of letters.
- $M(x, y)$ is the conditional frequency of observing the word y given that you observed x immediately before.

Setting up the problem

- Students formalized the problem as one in which we need to find:

$$f : \{\text{Message symbols}\} \to \{\text{Traditional alphabet}\}$$

- The number of possibilities for f is enormous.
- To judge how plausible was a particular f they introduce a linguistic model.
- They took War and Peace and estimated the conditional frequency of letters.
- $M(x, y)$ is the conditional frequency of observing the word y given that you observed x immediately before.

Setting up the problem

- Students formalized the problem as one in which we need to find:

$$f : \{\text{Message symbols}\} \to \{\text{Traditional alphabet}\}$$

- The number of possibilities for f is enormous.
- To judge how plausible was a particular f they introduce a linguistic model.
- They took War and Peace and estimated the conditional frequency of letters.
- $M(x, y)$ is the conditional frequency of observing the word y given that you observed x immediately before.

Setting up the problem

- Students formalized the problem as one in which we need to find:

$$f : \{\text{Message symbols}\} \to \{\text{Traditional alphabet}\}$$

- The number of possibilities for f is enormous.
- To judge how plausible was a particular f they introduce a linguistic model.
- They took War and Peace and estimated the conditional frequency of letters.
- $M(x, y)$ is the conditional frequency of observing the word y given that you observed x immediately before.

Solving the problem

- Solve:

$$\max_{f \in \mathfrak{F}} \prod M(f(s_i), f(s_{i+1})) \tag{1}$$

- This one is hard: use Metropolis Hasting algorithm.
- Check that your solution is plausible.

Solving the problem

- Solve:

$$\max_{f \in \mathfrak{F}} \prod M(f(s_i), f(s_{i+1})) \tag{1}$$

- This one is hard: use Metropolis Hasting algorithm.
- Check that your solution is plausible.

Solving the problem

- Solve:

$$\max_{f \in \mathfrak{F}} \prod M(f(s_i), f(s_{i+1})) \tag{1}$$

- This one is hard: use Metropolis Hasting algorithm.
- Check that your solution is plausible.

Validation

- Mix-up all letters in Shakespeare classic:

```
ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS
NOBLER IN THE MIND TO SUFFER THE SLINGS AND ARROWS OF OUTRAGEOUS
FORTUNE OR TO TAKE ARMS AGAINST A SEA OF TROUBLES AND BY OPPOSING END
```

Validation

- After running Metropolis Hasting n iterations:

```
 100 ER ENOHDLAE OHDLO UOZEOUNORU O UOZEO HD OITO HEOQSET IUROFHE HENO ITORUZAEN
 200 ES ELOHRNDE OHRNO UOVEOULOSU O UOVEO HR OITO HEOQAET IUSOPHE HELO ITOSUVDEL
 300 ES ELOHANDE OHANO UOVEOULOSU O UOVEO HA OITO HEOQRET IUSOFHE HELO ITOSUVDEL
 400 ES ELOHINME OHINO UOVEOULOSU O UOVEO HI OATO HEOQRET AUSOWHE HELO ATOSUVMEL
 500 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 600 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
 900 ES ELOHANME OHANO UODEOULOSU O UODEO HA OITO HEOQRET IUSOWHE HELO ITOSUDMEL
1000 IS ILOHANMI OHANO RODIORLOSR O RODIO HA OETO HIOQUIT ERSOWHI HILO ETOSRDMIL
1100 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1200 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1300 ISTILOHARMITOHAROT ODIO LOS TOT ODIOTHATOENOTHIOQUINTE SOWHITHILOTENOS DMIL
1400 ISTILOHAMRITOHAMOT OFIO LOS TOT OFIOTHATOENOTHIOQUINTE SOWHITHILOTENOS FRIL
1600 ESTEL HAMRET HAM TO CE OL SOT TO CE THAT IN THE QUENTIOS WHETHEL TIN SOCREL
1700 ESTEL HAMRET HAM TO BE OL SOT TO BE THAT IN THE QUENTIOS WHETHEL TIN SOBREL
1800 ESTER HAMLET HAM TO BE OR SOT TO BE THAT IN THE QUENTIOS WHETHER TIN SOBLER
1900 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
2000 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
```

Output

- The decoded message:

```
  to bat-rb. con todo mi respeto. i was sitting down playing chess with
danny de emf and boxer de el centro was sitting next to us. boxer was
making loud and loud voices so i tell him por favor can you kick back
homie cause im playing chess a minute later the vato starts back up again
so this time i tell him con respecto homie can you kick back.  the vato
stop for a minute and he starts up again so i tell him check this out shut
the f**k up cause im tired of your voice and if you got a problem with it
we can go to celda and handle it. i really felt disrespected thats why i
told him. anyways after i tell him that the next thing I know that vato
slashes me and leaves. dy the time i figure im hit i try to get away but
the c.o. is walking in my direction and he gets me right dy a celda. so i
go to the hole. when im in the hole my home boys hit doxer so now "b" is
also in the hole. while im in the hole im getting schoold wrong and
```

Contents

News-Based Economic Policy Uncertainty Index

- Developing news-based Economic Policy Uncertainty index with unsupervised machine learning: Azqueta-Gavaldón (2017)
- The EPU index objective is to capture snapshots of the unpredictability and incertitude of the Economy.
- The EPU index is constructed from frequencies of newspaper articles that contain a trio of terms pertaining to the economy, uncertainty and policy-related matters.

News-Based Economic Policy Uncertainty Index

- Developing news-based Economic Policy Uncertainty index with unsupervised machine learning: Azqueta-Gavaldón (2017)
- The EPU index objective is to capture snapshots of the unpredictability and incertitude of the Economy.
- The EPU index is constructed from frequencies of newspaper articles that contain a trio of terms pertaining to the economy, uncertainty and policy-related matters.

News-Based Economic Policy Uncertainty Index

- Developing news-based Economic Policy Uncertainty index with unsupervised machine learning: Azqueta-Gavaldón (2017)
- The EPU index objective is to capture snapshots of the unpredictability and incertitude of the Economy.
- The EPU index is constructed from frequencies of newspaper articles that contain a trio of terms pertaining to the economy, uncertainty and policy-related matters.

News-Based Economic Policy Uncertainty Index

- The index requires a meticulous manual process was needed. The index creation engages several research assistants to manually select those articles describing EPU from a pool of 12,000 articles containing the words economy and uncertainty.
- The paper proposes creating a news-based Economic Policy Uncertainty (EPU) index by employing an unsupervised algorithm able to deduce the subject of each article without the need for pre-labeled data.

News-Based Economic Policy Uncertainty Index

- The index requires a meticulous manual process was needed. The index creation engages several research assistants to manually select those articles describing EPU from a pool of 12,000 articles containing the words economy and uncertainty.
- The paper proposes creating a news-based Economic Policy Uncertainty (EPU) index by employing an unsupervised algorithm able to deduce the subject of each article without the need for pre-labeled data.

EPU index sample

- The monthly EPU index for the United States relies on 10 leading newspapers (*USA Today*, *Los Angeles Times*, *Wall Street Journal*, ...)
- Articles taken into account contained the following trio of terms:
  1. "uncertainty" or "uncertain"
  2. "economic" or "economy"
  3. "congress," "deficit," "Federal Reserve," "legislation," "regulation," or "White House"
- This process is done manually and it's classified into categories such as Fiscal Policy, Monetary Policy, Healthcare, National Security, ... for further disaggregated analysis of the index.
- The construction of the index since 1985 to 2017 took 2 years

EPU index sample

- The monthly EPU index for the United States relies on 10 leading newspapers (*USA Today, Los Angeles Times, Wall Street Journal, ...*)
- Articles taken into account contained the following trio of terms:
  1. "uncertainty" or "uncertain"
  2. "economic" or "economy"
  3. "congress," "deficit," "Federal Reserve," "legislation," "regulation," or "White House"
- This process is done manually and it's classified into categories such as Fiscal Policy, Monetary Policy, Healthcare, National Security, ... for further disaggregated analysis of the index.
- The construction of the index since 1985 to 2017 took 2 years

EPU index sample

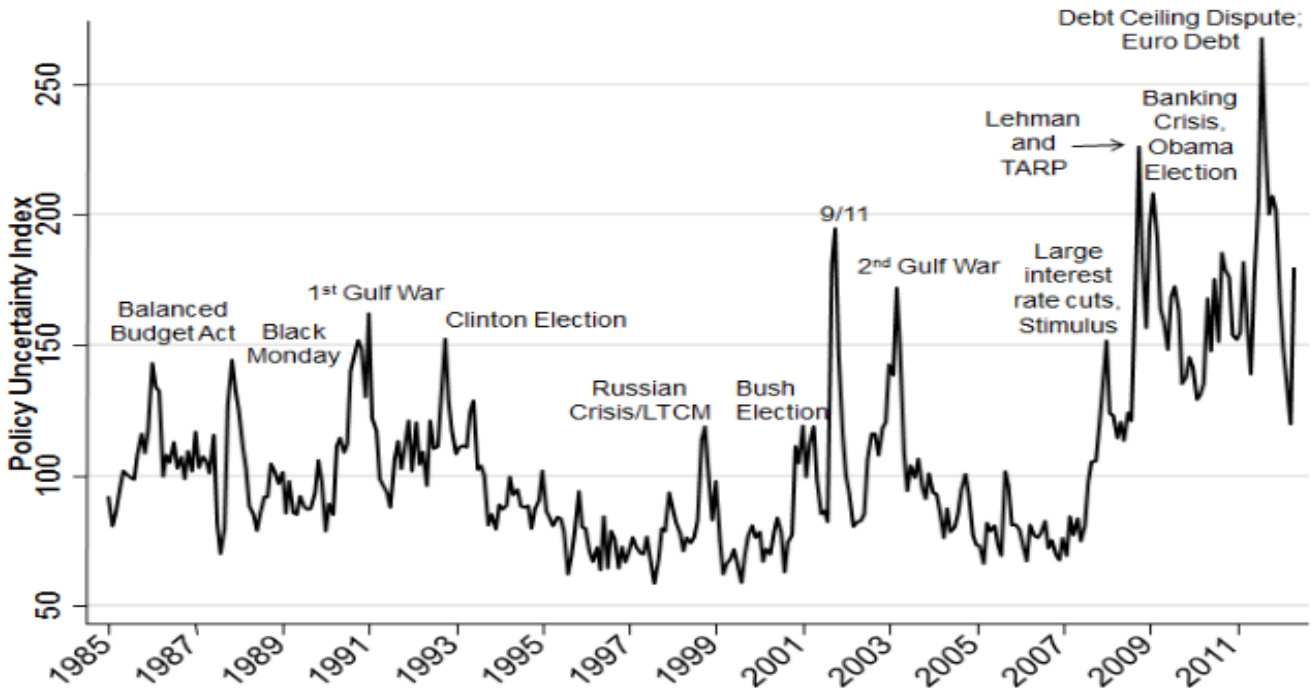
- The monthly EPU index for the United States relies on 10 leading newspapers (*USA Today, Los Angeles Times, Wall Street Journal, ...*)
- Articles taken into account contained the following trio of terms:
  1. "uncertainty" or "uncertain"
  2. "economic" or "economy"
  3. "congress," "deficit," "Federal Reserve," "legislation," "regulation," or "White House"
- This process is done manually and it's classified into categories such as Fiscal Policy, Monetary Policy, Healthcare, National Security, ... for further disaggregated analysis of the index.
- The construction of the index since 1985 to 2017 took 2 years

EPU index sample

- The monthly EPU index for the United States relies on 10 leading newspapers (*USA Today, Los Angeles Times, Wall Street Journal, ...*)
- Articles taken into account contained the following trio of terms:
  1. "uncertainty" or "uncertain"
  2. "economic" or "economy"
  3. "congress," "deficit," "Federal Reserve," "legislation," "regulation," or "White House"
- This process is done manually and it's classified into categories such as Fiscal Policy, Monetary Policy, Healthcare, National Security, ... for further disaggregated analysis of the index.
- The construction of the index since 1985 to 2017 took 2 years

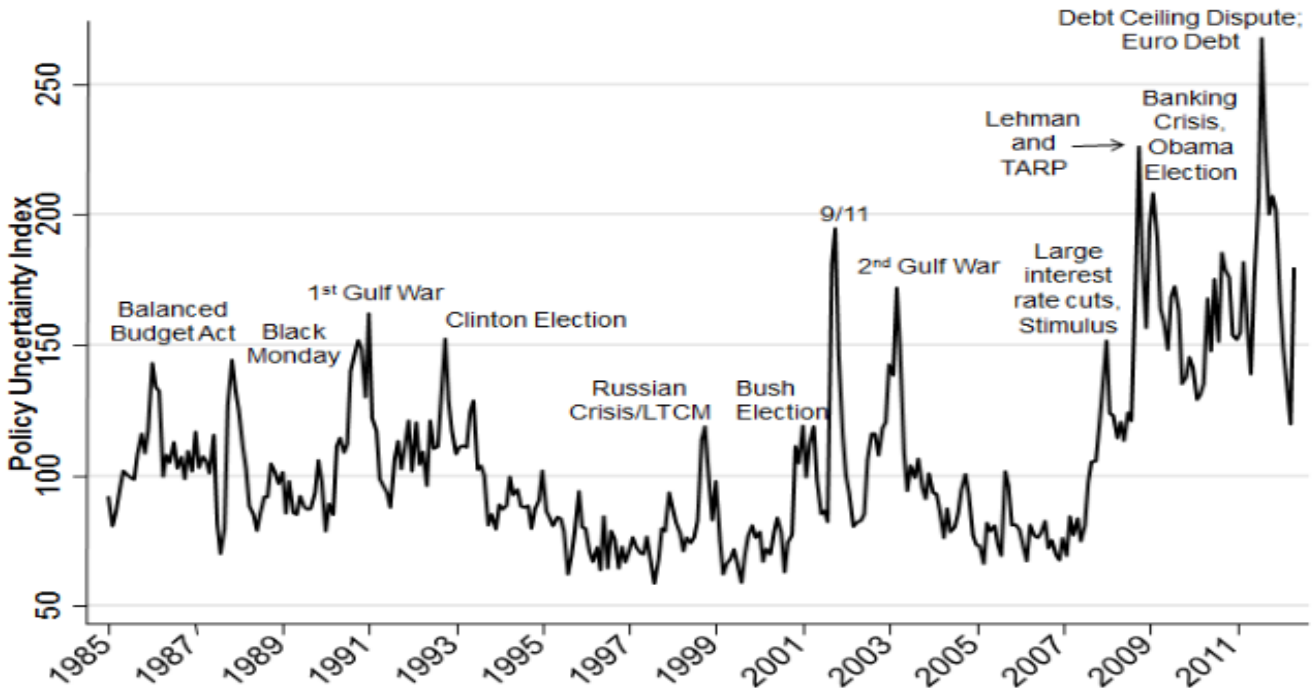
EPU index construction

- Each month the number of articles containing the terms showed above are count.
- Each article is labeled into one of 8 categories to create and index for specific topics.
- The series of number of articles is normalized to a mean of 100.



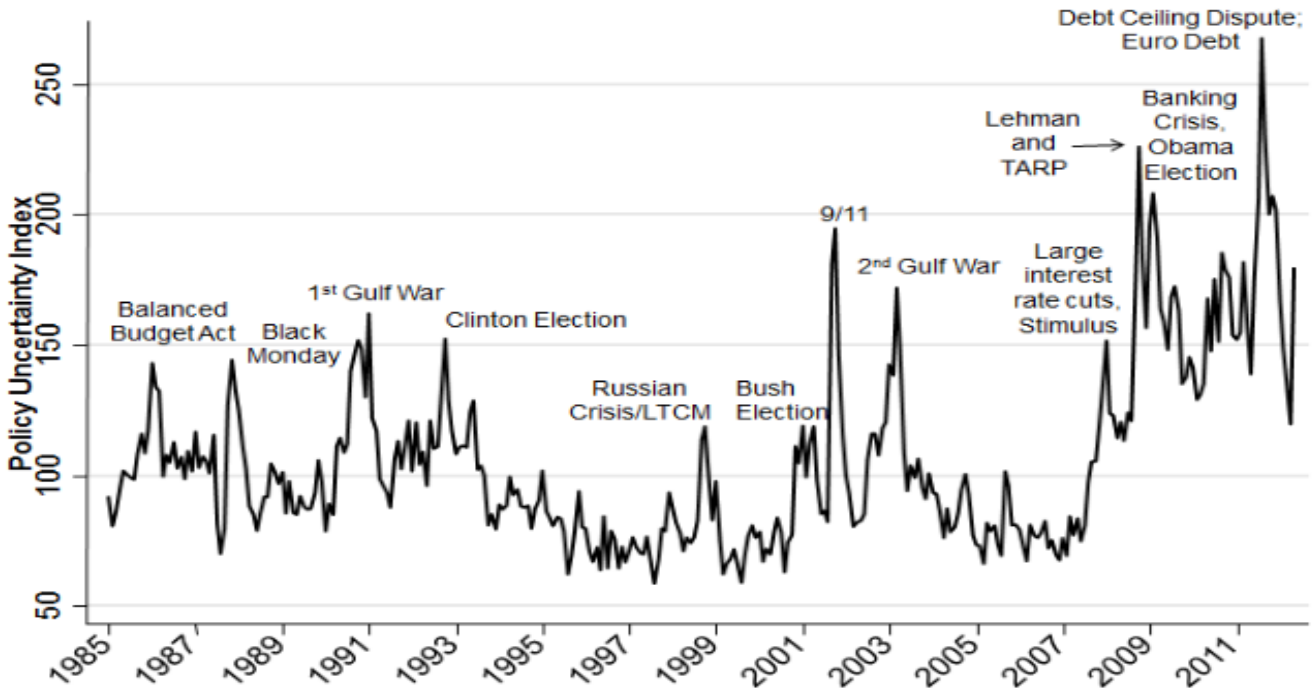
EPU index construction

- Each month the number of articles containing the terms showed above are count.
- Each article is labeled into one of 8 categories to create and index for specific topics.
- The series of number of articles is normalized to a mean of 100.



EPU index construction

- Each month the number of articles containing the terms showed above are count.
- Each article is labeled into one of 8 categories to create and index for specific topics.
- The series of number of articles is normalized to a mean of 100.



Machine Learning to recreate EPU index

- The author proposes the use of Latent Dirichlet Allocation (LDA), an unsupervised machine learning algorithm, to reduce the cost of estimation of the index.
- By making the machine generate the topics the labeling of articles is done automatically, instead of having people manually classifying them into categories.

Machine Learning to recreate EPU index

The new method proposed by the author consists of the next steps:

1. Download all the articles containing *economy* and *uncertainty* from USA Today, The New York Times and The Washington Post from 1989 to 2016 (40,454 articles)
2. Pre-process the articles using *stopwords*
3. Apply LDA and find the most likely value of topics K for this specific corpus
4. The documents are labeled as the topic to which they belong the most
5. Group the topics generated by the LDA into the 8 categories that compose de EPU index. This part is the most manual of the process, since the researcher has to interpret the topics made by the LDA.
6. Normalize the time series by dividing each raw time-series by the total number of articles containing the word today each month (the proxy for the total number of articles)
7. Finally standarize to mean 100 and one standard deviation the time series.

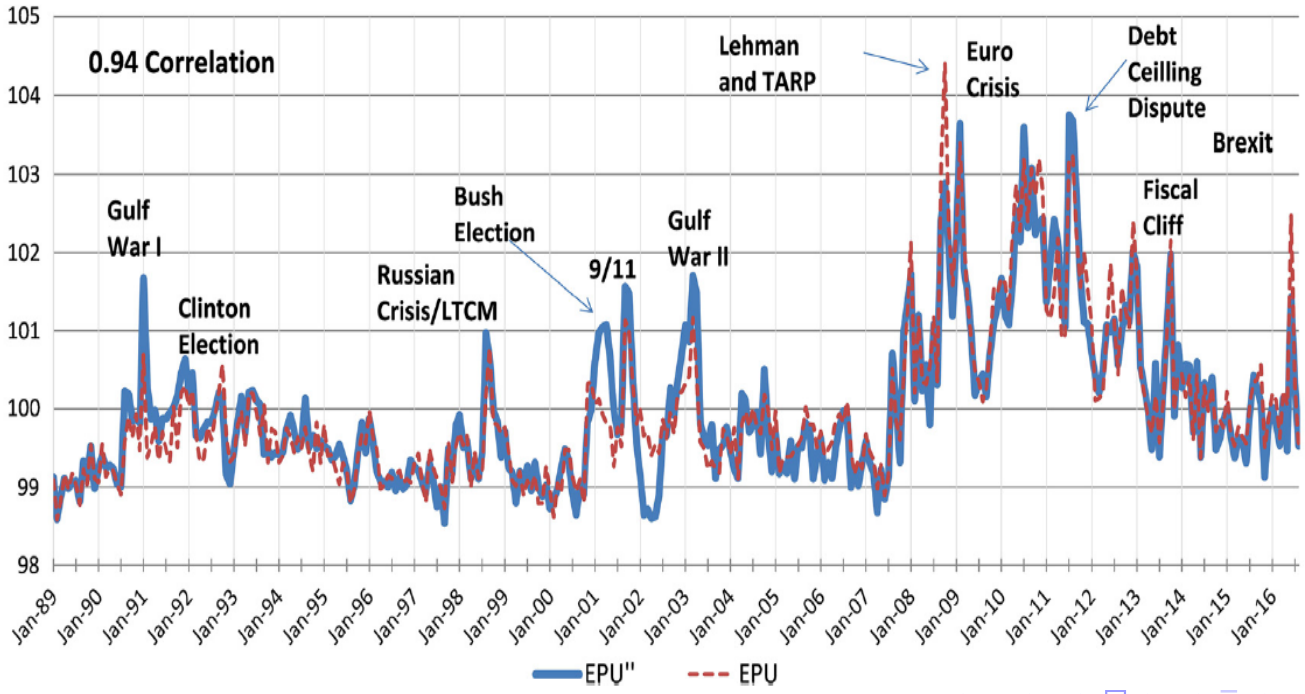
Machine Learning to recreate EPU index

The new method proposed by the author consists of the next steps:

1. Download all the articles containing *economy* and *uncertainty* from USA Today, The New York Times and The Washington Post from 1989 to 2016 (40,454 articles)
2. Pre-process the articles using *stopwords*
3. Apply LDA and find the most likely value of topics K for this specific corpus
4. The documents are labeled as the topic to which they belong the most
5. Group the topics generated by the LDA into the 8 categories that compose de EPU index. This part is the most manual of the process, since the researcher has to interpret the topics made by the LDA.
6. Normalize the time series by dividing each raw time-series by the total number of articles containing the word today each month (the proxy for the total number of articles)
7. Finally standarize to mean 100 and one standard deviation the time series.

Results

- 0.94 correlation between the original index and the proposed
- 0.88 correlation in the cyclical component
- 0.99 correlation in the trend component
- The construction of the index took less than a week



The Effects of the Content of FOMC Communications on US Treasury Rates: Rohlfs et al. (2016)

- The paper measures the effects of Federal Open Market Committee (FOMC) text content on the direction of short- and medium-term interest rate movements.
- FOMC meets roughly eight times a year and releases a statement (information about interest rates, the money supply, inflation, unemployment, and economic growth) after each meeting.
- Literature suggests that:
  - Equity and interest rate markets tend to move when FOMC communications are released.
  - Policy actions do not explain these responses.
  - The directions of market movements coincide with a human-coded measure of the sentiment expressed in the texts.

The Effects of the Content of FOMC Communications on US Treasury Rates: Rohlfs et al. (2016)

- The paper measures the effects of Federal Open Market Committee (FOMC) text content on the direction of short- and medium-term interest rate movements.
- FOMC meets roughly eight times a year and releases a statement (information about interest rates, the money supply, inflation, unemployment, and economic growth) after each meeting.
- Literature suggests that:
  - Equity and interest rate markets tend to move when FOMC communications are released.
  - Policy actions do not explain these responses.
  - The directions of market movements coincide with a human-coded measure of the sentiment expressed in the texts.

The Effects of the Content of FOMC Communications on US Treasury Rates: Rohlfs et al. (2016)

- The paper measures the effects of Federal Open Market Committee (FOMC) text content on the direction of short- and medium-term interest rate movements.
- FOMC meets roughly eight times a year and releases a statement (information about interest rates, the money supply, inflation, unemployment, and economic growth) after each meeting.
- Literature suggests that:
  - Equity and interest rate markets tend to move when FOMC communications are released.
  - Policy actions do not explain these responses.
  - The directions of market movements coincide with a human-coded measure of the sentiment expressed in the texts.

The Effects of the Content of FOMC Communications on US Treasury Rates: Rohlfs et al. (2016)

- The paper measures the effects of Federal Open Market Committee (FOMC) text content on the direction of short- and medium-term interest rate movements.
- FOMC meets roughly eight times a year and releases a statement (information about interest rates, the money supply, inflation, unemployment, and economic growth) after each meeting.
- Literature suggests that:
  - Equity and interest rate markets tend to move when FOMC communications are released.
  - Policy actions do not explain these responses.
  - The directions of market movements coincide with a human-coded measure of the sentiment expressed in the texts.

The Effects of the Content of FOMC Communications on US Treasury Rates: Rohlfs et al. (2016)

- The paper measures the effects of Federal Open Market Committee (FOMC) text content on the direction of short- and medium-term interest rate movements.
- FOMC meets roughly eight times a year and releases a statement (information about interest rates, the money supply, inflation, unemployment, and economic growth) after each meeting.
- Literature suggests that:
  - Equity and interest rate markets tend to move when FOMC communications are released.
  - Policy actions do not explain these responses.
  - The directions of market movements coincide with a human-coded measure of the sentiment expressed in the texts.

Methodology

- The authors use MedLDA (Maximum Margin Supervised Topic Models).
- They address possible bias from one important set of omitted variables—releases of macroeconomic data- by estimating specifications in which they control for those factors separately and predict whether interest rates moved more or less than would be expected based upon the latest data on the macroeconomic environment (Consumer Price Index, Unemployment, US Gross Domestic Product).

Methodology

- The authors use MedLDA (Maximum Margin Supervised Topic Models).
- They address possible bias from one important set of omitted variables—releases of macroeconomic data- by estimating specifications in which they control for those factors separately and predict whether interest rates moved more or less than would be expected based upon the latest data on the macroeconomic environment (Consumer Price Index, Unemployment, US Gross Domestic Product).

Methodology

- They control for time-specific factors like day-of-week effects and time trends using only observations from non-FOMC dates.
- The study splits the sample into training and test sets in order to limit over fitting in the predicted values.
- Sample: observations from meetings from May 1999 to May 2016.

Methodology

- They control for time-specific factors like day-of-week effects and time trends using only observations from non-FOMC dates.
- The study splits the sample into training and test sets in order to limit over fitting in the predicted values.
- Sample: observations from meetings from May 1999 to May 2016.

Methodology

- They control for time-specific factors like day-of-week effects and time trends using only observations from non-FOMC dates.
- The study splits the sample into training and test sets in order to limit over fitting in the predicted values.
- Sample: observations from meetings from May 1999 to May 2016.

Methodology: Pre-processing

- FOMC are highly repetitive containing nearly identical sentences and sentence structures from previous meetings.
- Pre-processing of text:
  - Non-alphabetic characters are removed.
  - The texts are converted to lower case.
  - Each document is separated into a bag of words.
  - Common words and stop words are deleted.
  - Words are stemmed using the Porter stemming algorithm (a process for removing the commoner morphological and inflexional endings from words in English).
  - One letter words are dropped.

Methodology: Pre-processing

- FOMC are highly repetitive containing nearly identical sentences and sentence structures from previous meetings.
- Pre-processing of text:
  - Non-alphabetic characters are removed.
  - The texts are converted to lower case.
  - Each document is separated into a bag of words.
  - Common words and stop words are deleted.
  - Words are stemmed using the Porter stemming algorithm (a process for removing the commoner morphological and inflexional endings from words in English).
  - One letter words are dropped.

Methodology: MedLDA

- MedLDA (Maximum Margin Supervised Topic Models) is a supervised-topic model that can be built for both regression and classification problems.
- The study built the MedLDA for classification task (response variable y has three discrete values $\{1, 0, -1\}$ denoting the movements of interest rates).

Methodology: MedLDA

- MedLDA (Maximum Margin Supervised Topic Models) is a supervised-topic model that can be built for both regression and classification problems.
- The study built the MedLDA for classification task (response variable y has three discrete values $\{1, 0, -1\}$ denoting the movements of interest rates).

Results

- In most cases, the best accuracy is given by K = 20.
- With no control variables → accuracy is 93% and 64% for the Target and Effective Federal Funds Rate.
- With control variables → similar but somewhat lower accuracy rates.

Table 1: Accuracy of Medlda Classifier after purging out of control for statements between 1999 and May, 2016 [K (topics) = 20]

Outcome variable	MedLDA			Baseline (Random Chance)[5]		
	None	Linear	Interactions	None	Linear	Interactions
Target Fed Funds Rate	0.9321	0.9160	0.8954	0.6849	0.6849	0.6849
Effective Fed Funds Rate	0.6421	0.4479	0.5112	0.4589	0.4658	0.4658
Median Treasury Rate	0.4209	0.3803	0.4012	0.4589	0.4247	0.4247
Average Treasury Rate	0.3803	0.4611	0.3924	0.4726	0.4041	0.4041

- MedLDA model can effectively associate the text contents of the meetings with the movements in the rate.