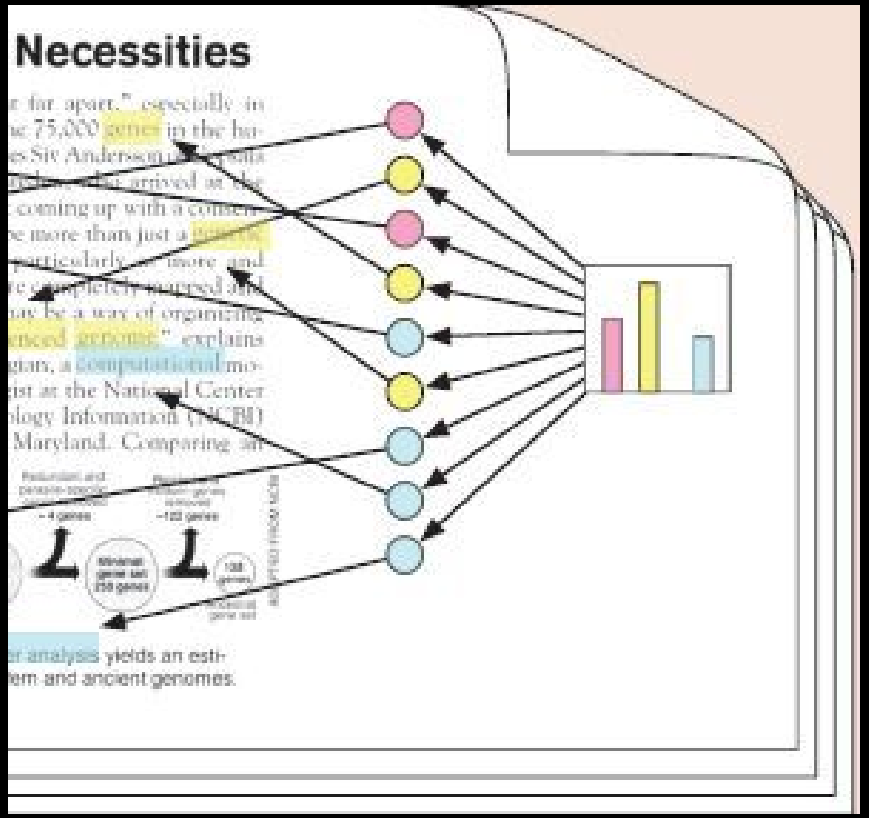
the Declassification Engine

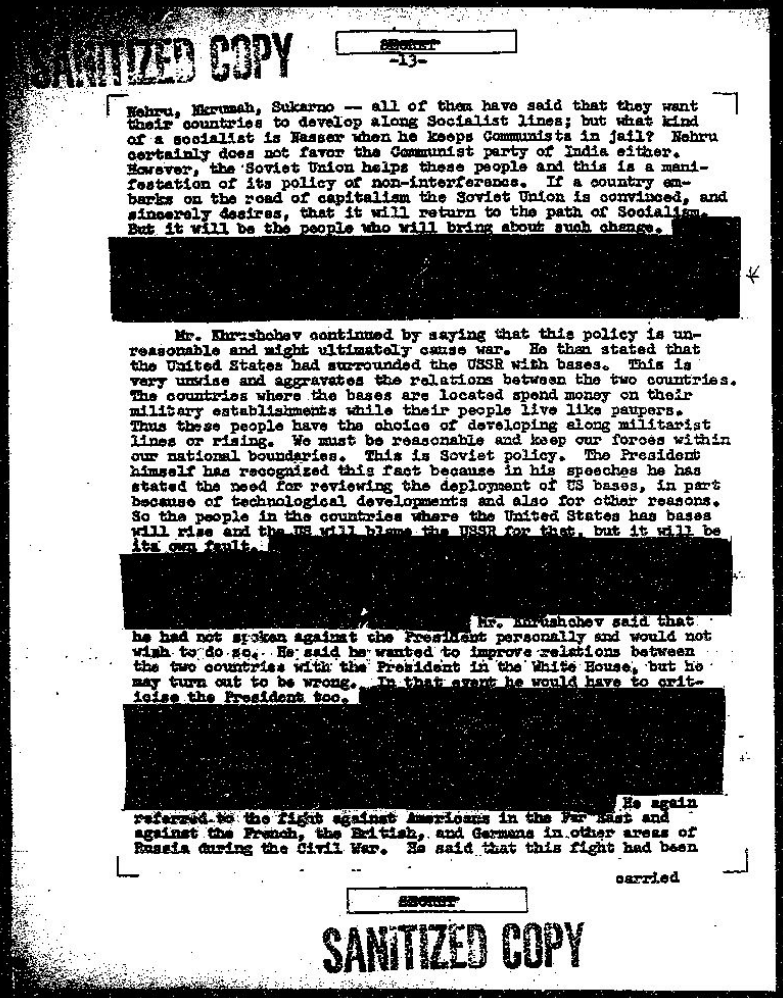
Sophie Chou
2014.05.06

Contents

Topic Modelling on & around Redactions



Sophie Chou
2014.05.06


SANITIZED COPY

SECRET

-13-

Nehru, Nkrumah, Sukarno — all of them have said that they want their countries to develop along Socialist lines; but what kind of a socialist is Nasser when he keeps Communists in jail? Nehru certainly does not favor the Communist party of India either. However, the Soviet Union helps these people and this is a manifestation of its policy of non-interference. If a country embarks on the road of capitalism the Soviet Union is convinced, and sincerely desires, that it will return to the path of Socialism. But it will be the people who will bring about such change.

[redacted]

Mr. Khrushchev continued by saying that this policy is unreasonable and might ultimately cause war. He then stated that the United States had surrounded the USSR with bases. This is very unwise and aggravates the relations between the two countries. The countries where the bases are located spend money on their military establishments while their people live like paupers. Thus these people have the choice of developing along militarist lines or rising. We must be reasonable and keep our forces within our national boundaries. This is Soviet policy. The President himself has recognized this fact because in his speeches he has stated the need for reviewing the deployment of US bases, in part because of technological developments and also for other reasons. So the people in the countries where the United States has bases will rise and the US will blame the USSR for that, but it will be its own fault.

[redacted]

Mr. Khrushchev said that he had not spoken against the President personally and would not wish to do so. He said he wanted to improve relations between the two countries with the President in the White House, but he may turn out to be wrong. In that event he would have to criticise the President too.

[redacted]

He again referred to the fight against Americans in the Far East and against the French, the British, and Germans in other areas of Russia during the Civil War. He said that this fight had been carried

SECRET

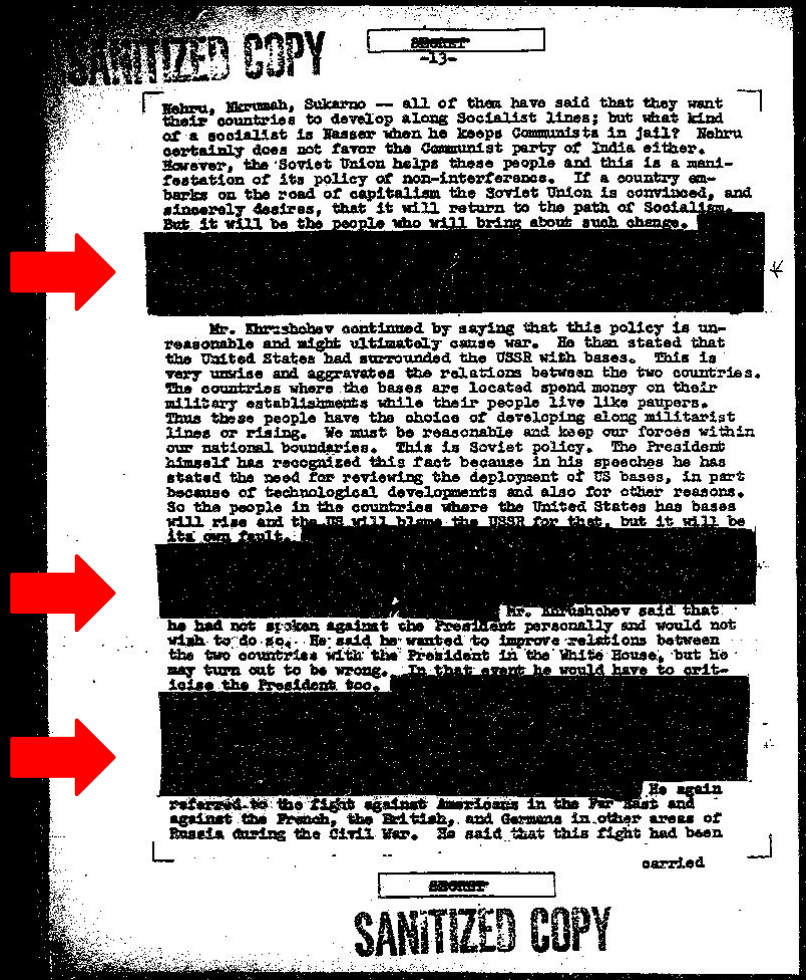
SANITIZED COPY

Recap: Detecting Redactions with Computer Vision

Previously...

- ★ How can we find redactions in microfilmed images of historical documents?
- ★ Documents are noisy, non-uniform, and of poor quality.
- ★ Image-based rather than text-based approach

Focus on dark redactions & the Kennedy archives (100 docs)



SANITIZED COPY

SECRET

-13-

Nehru, Nkrumah, Sukarno -- all of them have said that they want their countries to develop along Socialist lines; but what kind of a socialist is Nasser when he keeps Communists in jail? Nehru certainly does not favor the Communist party of India either. However, the Soviet Union helps these people and this is a manifestation of its policy of non-interference. If a country embarks on the road of capitalism the Soviet Union is convinced, and sincerely desires, that it will return to the path of Socialism. But it will be the people who will bring about such change.

[redacted]

Mr. Khrushchev continued by saying that this policy is unreasonable and might ultimately cause war. He then stated that the United States had surrounded the USSR with bases. This is very unwise and aggravates the relations between the two countries. The countries where the bases are located spend money on their military establishments while their people live like paupers. Thus these people have the choice of developing along militarist lines or rising. We must be reasonable and keep our forces within our national boundaries. This is Soviet policy. The President himself has recognized this fact because in his speeches he has stated the need for reviewing the deployment of US bases, in part because of technological developments and also for other reasons. So the people in the countries where the United States has bases will rise and the US will blame the USSR for that, but it will be its own fault.

[redacted]

Mr. Khrushchev said that he had not spoken against the President personally and would not wish to do so. He said he wanted to improve relations between the two countries with the President in the White House, but he may turn out to be wrong. In that event he would have to criticise the President too.

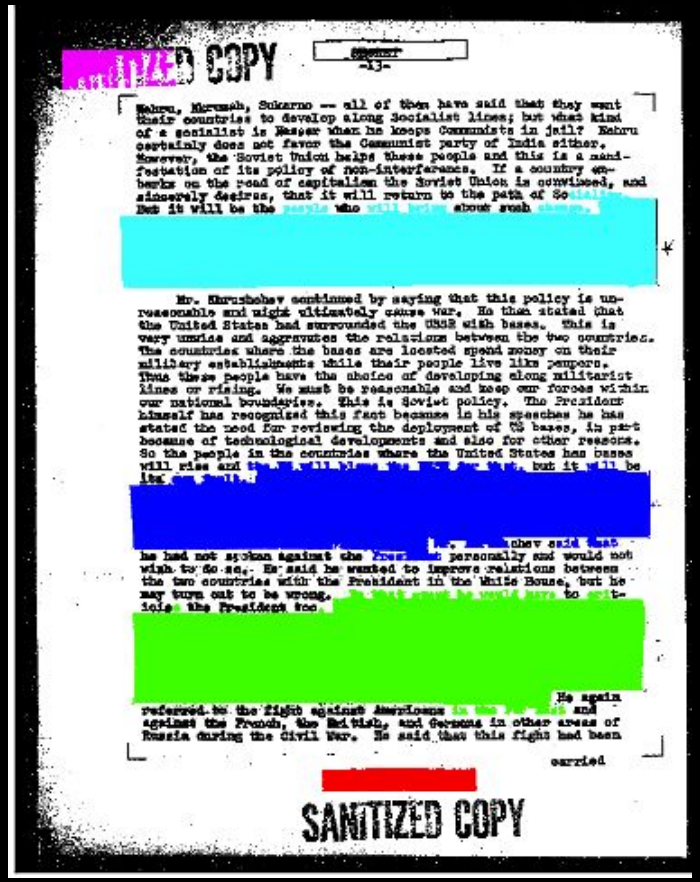
[redacted]

He again referred to the fight against Americans in the Far East and against the French, the British, and Germans in other areas of Russia during the Civil War. He said that this fight had been carried

SECRET

SANITIZED COPY

Difficulties



SECRET
-13-

Nehru, Nkrumah, Sukarno — all of them have said that they want their countries to develop along Socialist lines; but what kind of a socialist is Nasser when he keeps Communists in jail? Nehru certainly does not favor the Communist party of India either. However, the Soviet Union helps these people and this is a manifestation of its policy of non-interference. If a country embarks on the road of capitalism the Soviet Union is convinced, and sincerely desires, that it will return to the path of Socialism. But it will be the people who will bring about such change.

Mr. Khrushchev continued by saying that this policy is unreasonable and might ultimately cause war. He then stated that the United States had surrounded the USSR with bases. This is very unwise and aggravates the relations between the two countries. The countries where the bases are located spend money on their military establishments while their people live like paupers. Thus these people have the choice of developing along militarist lines or rising. We must be reasonable and keep our forces within our national boundaries. This is Soviet policy. The President himself has recognized this fact because in his speeches he has stated the need for reviewing the deployment of US bases, in part because of technological developments and also for other reasons. So the people in the countries where the United States has bases will rise and the US will blame the USSR for that, but it will be its own fault.

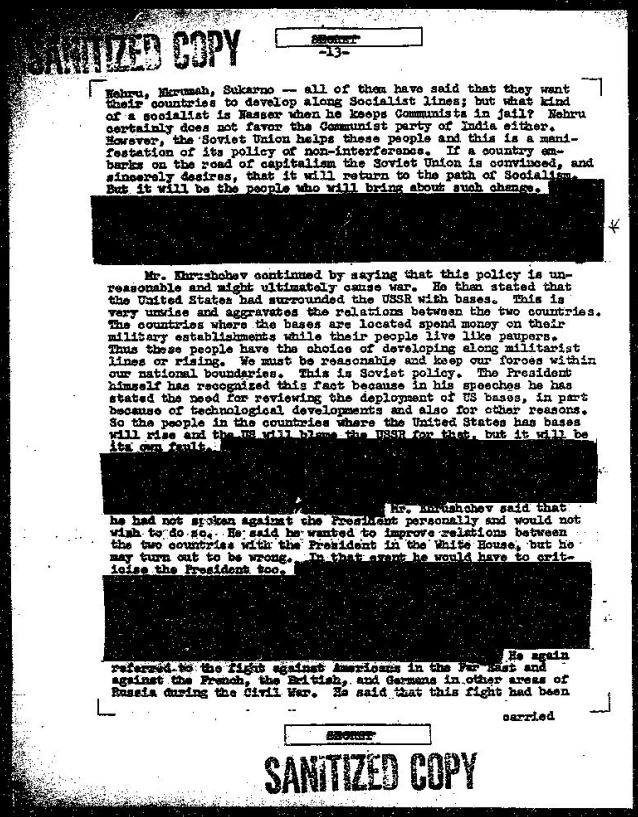
Mr. Khrushchev said that he had not spoken against the President personally and would not wish to do so. He said he wanted to improve relations between the two countries with the President in the White House, but he may turn out to be wrong. In that event he would have to criticise the President too.

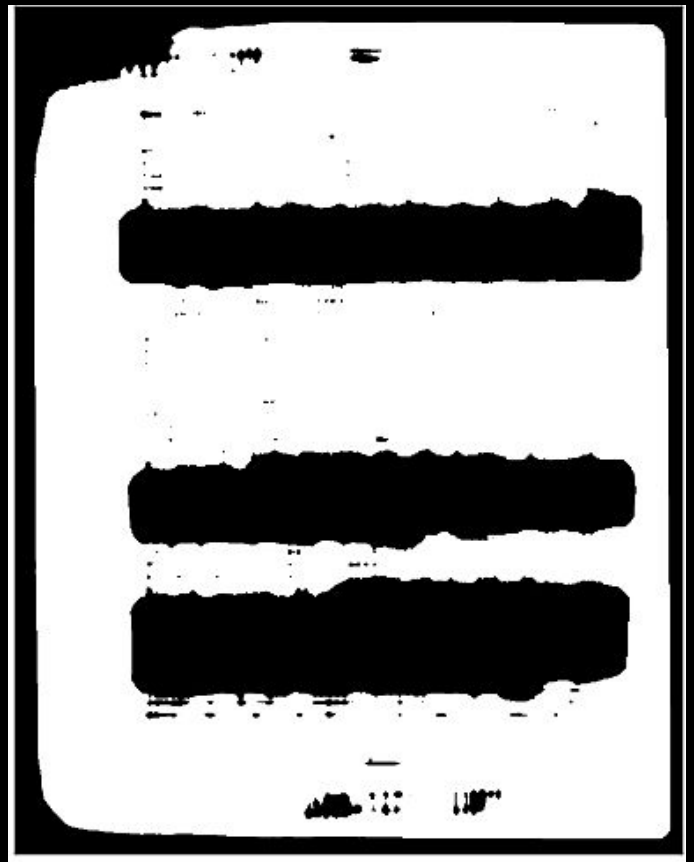
He again referred to the fight against Americans in the Far East and against the French, the British, and Germans in other areas of Russia during the Civil War. He said that this fight had been carried

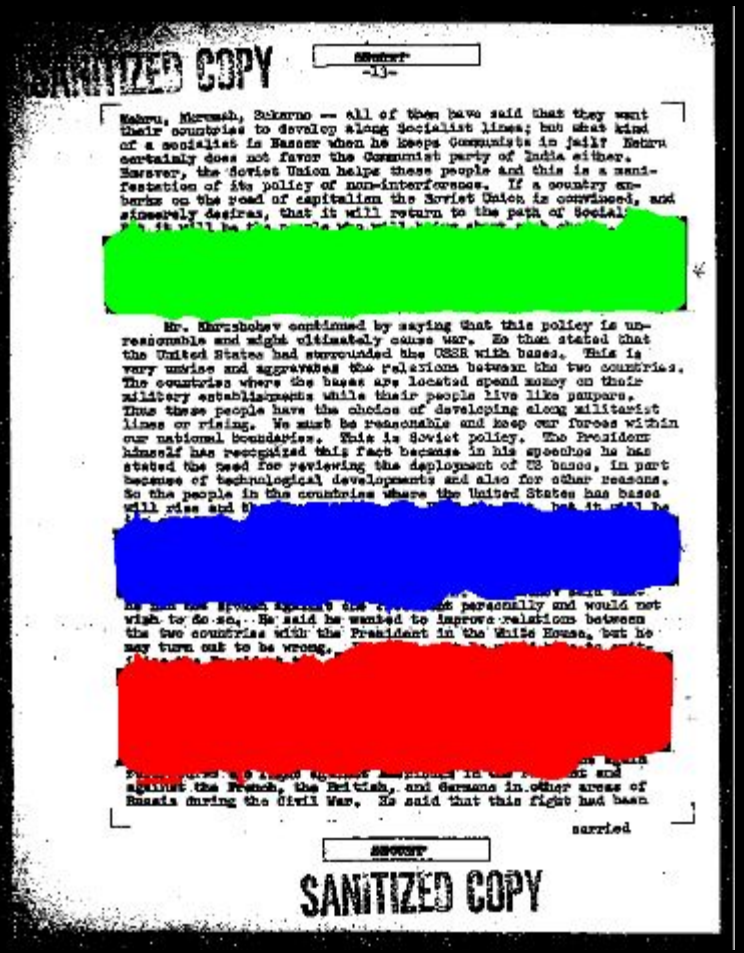
SECRET

SANITIZED COPY

Median filtering







Median Filter

1. apply median filter of k-size aperture on documents
2. Detect remaining blobs, limited by parameters

Results?

73% accuracy,
63% F-Score

Boolean Predictions

- Evaluated on testing, mean of 98% correct
- Future work

Last semester's presentation is available here.

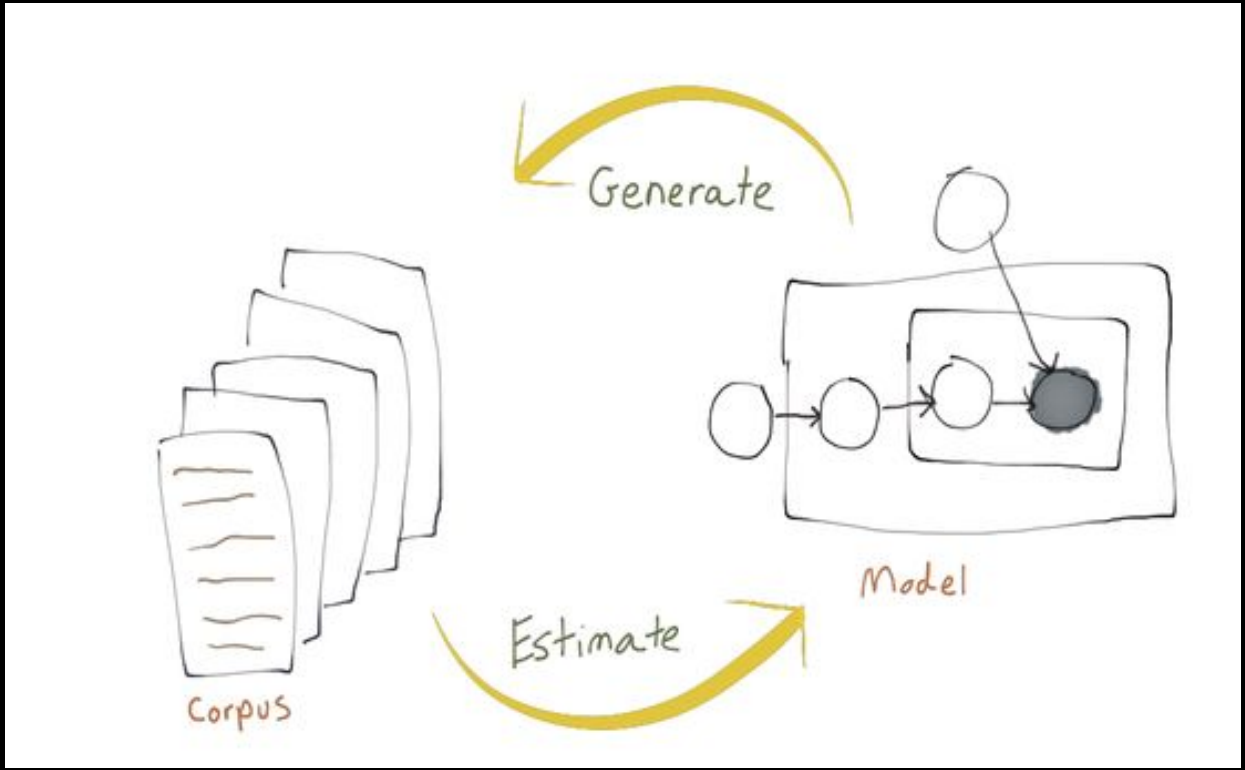
The code is available here.

Topic Modelling on FRUS Documents

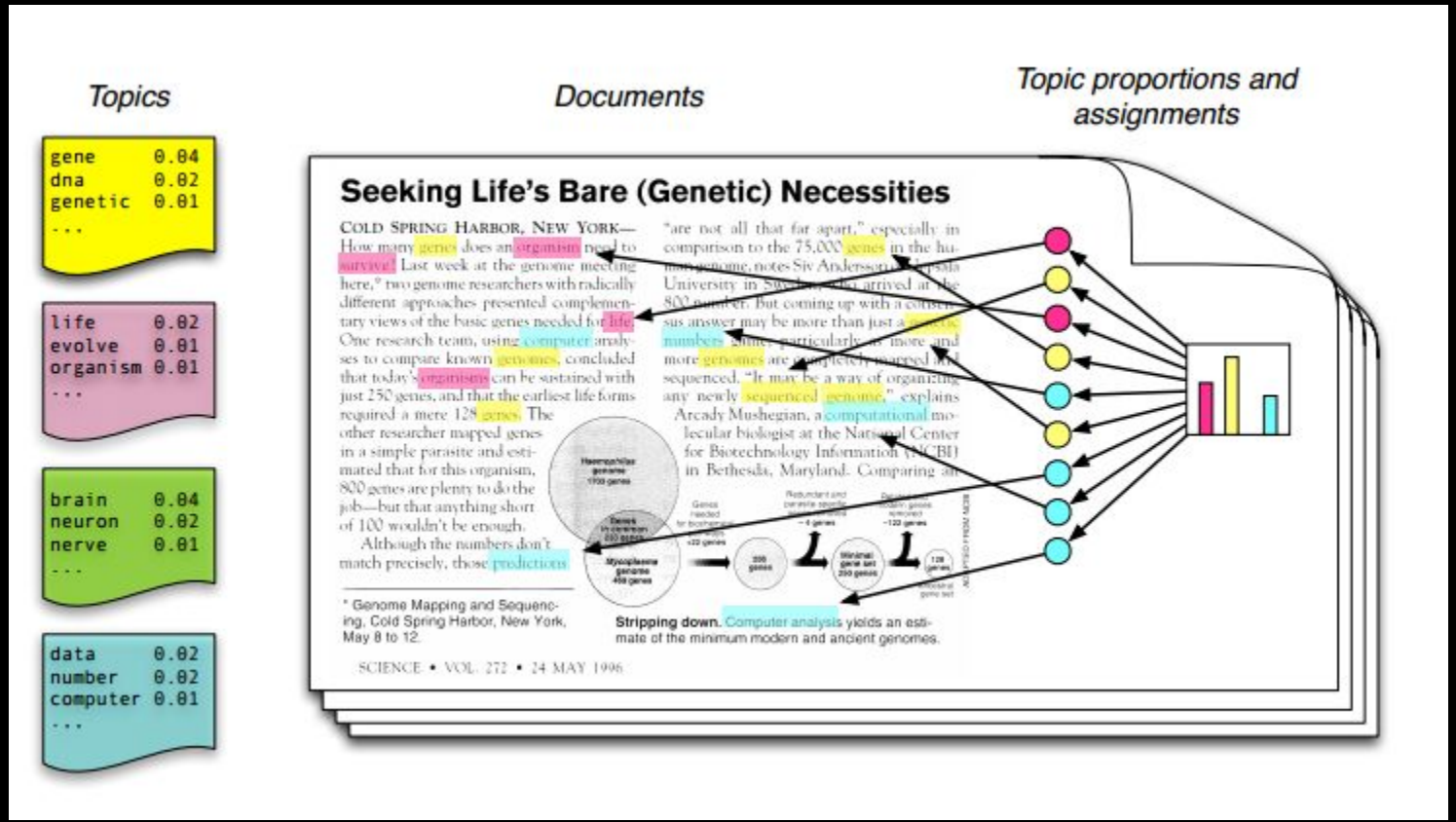
The Goals of Topic Modelling

- ★ Discover the hidden themes in a collection of documents (topics)
- ★ Annotate the documents according to those themes
- ★ Applications : use annotations to organize, summarize, and search the texts

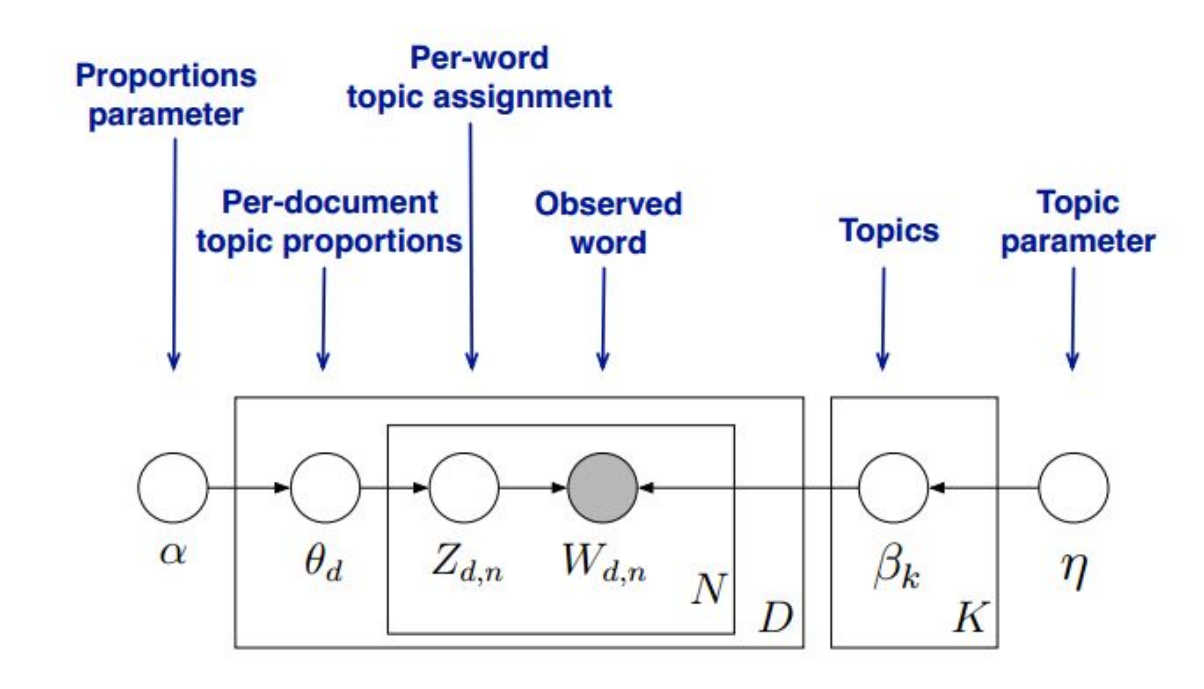
Modelling Basics



The Generative Model



Latent Dirichlet Allocation



Computation

- ★ We want to compute the *posterior*
- ★ (recall that the posterior p(theta | x) is the probability of latent variable theta given evidence X)
- ★ Estimate p(topics, proportions, assignments | documents)

Text Predictions

on Document Pairs
(Sasha Rush)

[redacted]

-3-

involved in operations of uncertain dimensions. This involvement is taking place in a political framework which is unclear to our allies, our adversaries, the people of Southeast Asia, the American public, and the Congress.

The greatest danger may lie in the profound confusion that exists among the 200 million Southeast Asians who are most directly concerned. Most young anti-Communist Vietnamese, Cambodians, Laos, Burmese, Malayans and Indonesians now tend to think of the U.S. in terms of the massive military supplies which we sent the French "colonialists" in the early 1950's, [redacted] and of our support for Ngric, Diem, Phoumi and Chiang Kai-shek, none of whom can be said to reflect the "New Frontier" in Asia.

It is essential that we present ourselves in a fresh and affirmative role. This requires a governmental decision, followed by a public statement understandable both

[redacted]

SECRET

-3-

involved in operations of uncertain dimensions. This involvement is taking place in a political framework which is unclear to our allies, our Adversaries, the people of Southeast Asia, the American public, and the Congress.

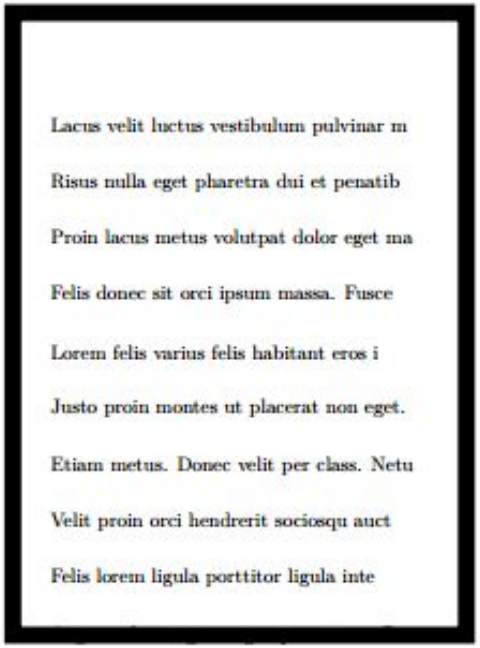
The greatest danger may lie in the profound confusion that exists among the 200 million Southeast Asians who are most directly concerned. Most young anti-Communist Vietnamese, Cambodians, Laos, Burmese, Malayans and Indonesians now tend to think of the U.S. in terms of the massive military supplies which we sent the French "colonialists" in the early 1950's, of our abortive efforts to build up a Nationalist Chinese force within the borders of independent Burma, of our CIA efforts to overthrow Sukarno in 1958, and of our support for Ngric, Diem, Phoumi and Chiang Kai-shek, none of whom can be said to reflect the "New Frontier" in Asia.

It is essential that we present ourselves in a fresh and affirmative role. This requires a governmental decision, followed by a public statement understandable both

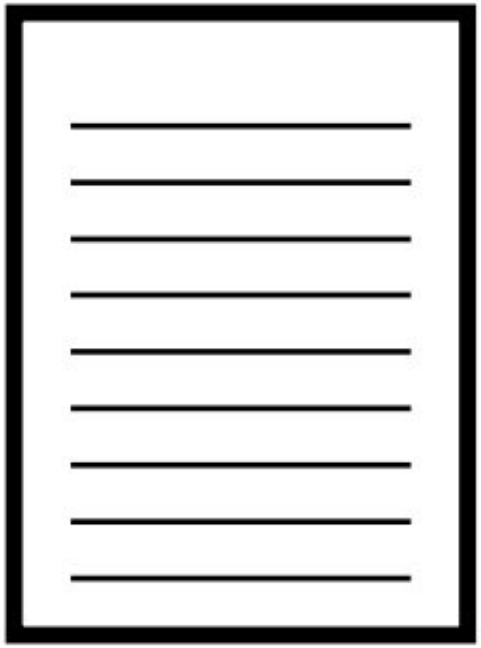
SECRET

Source Document

x

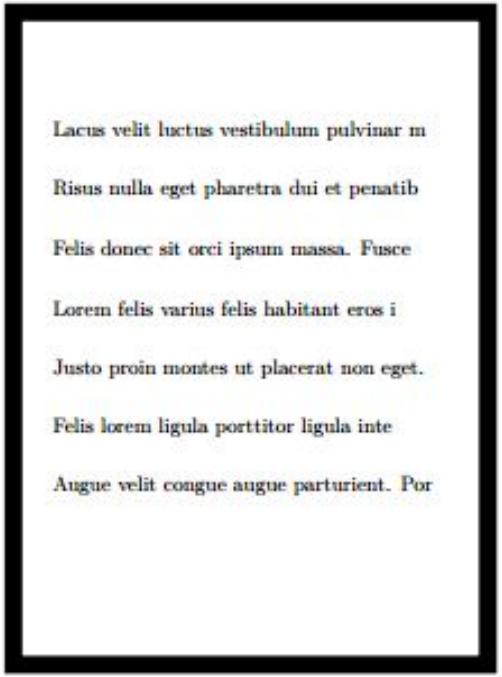


y

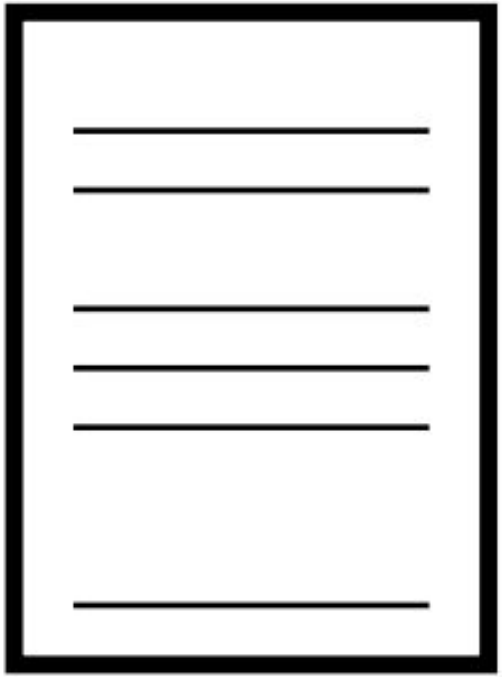


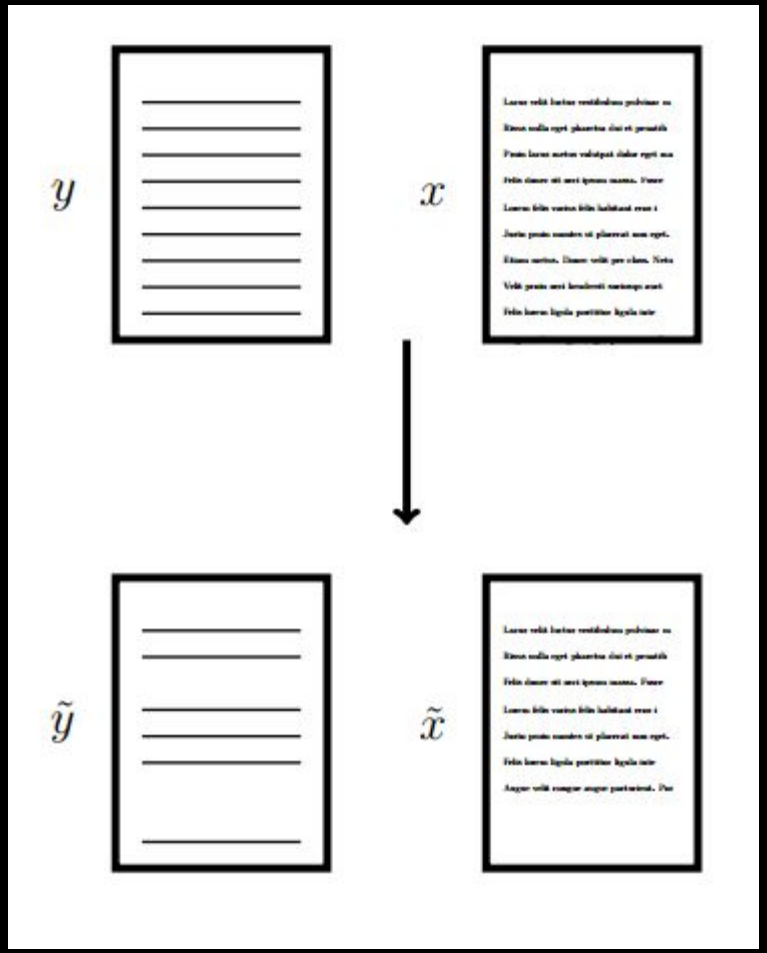
Observed Document

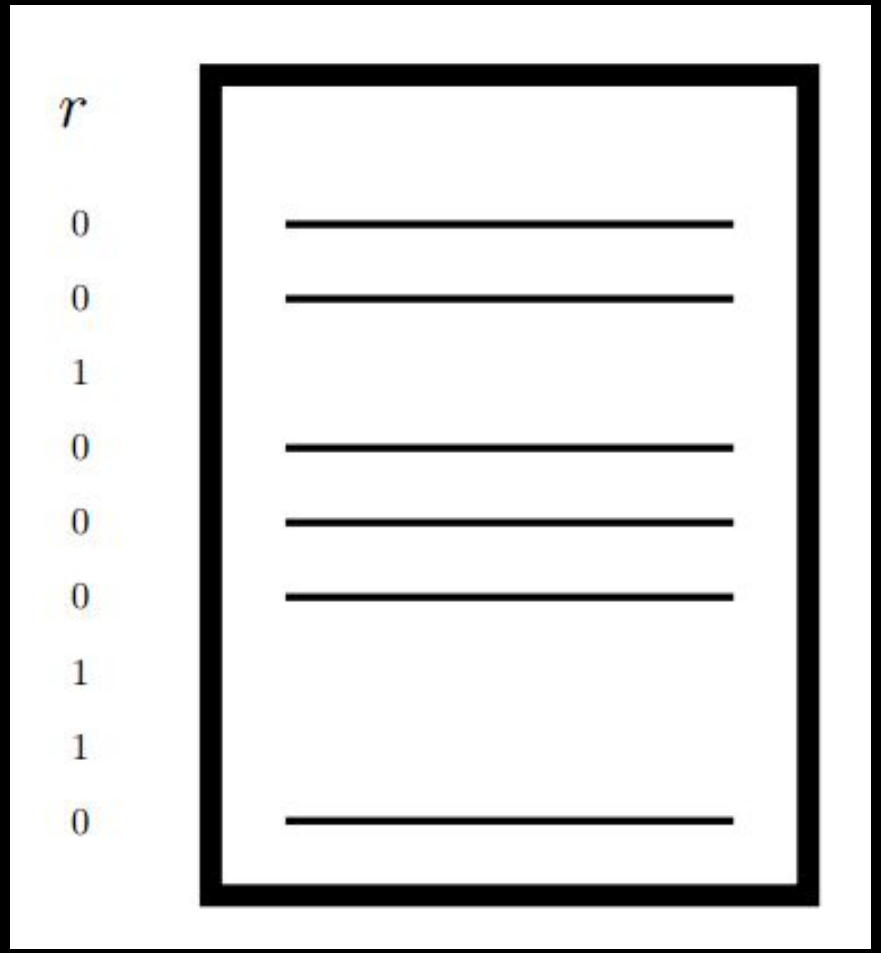
$\tilde{x}$



$\tilde{y}$




y
x
$\tilde{y}$
$\tilde{x}$


r
0
0
1
0
0
0
1
1
0

- $\tilde{x}^{(1)}, \tilde{y}^{(1)}$; the first document text and layout

$\tilde{y}^{(1)}$  $\tilde{x}^{(1)}$ 

- $\tilde{x}^{(2)}, \tilde{y}^{(2)}$; the second document text and layout

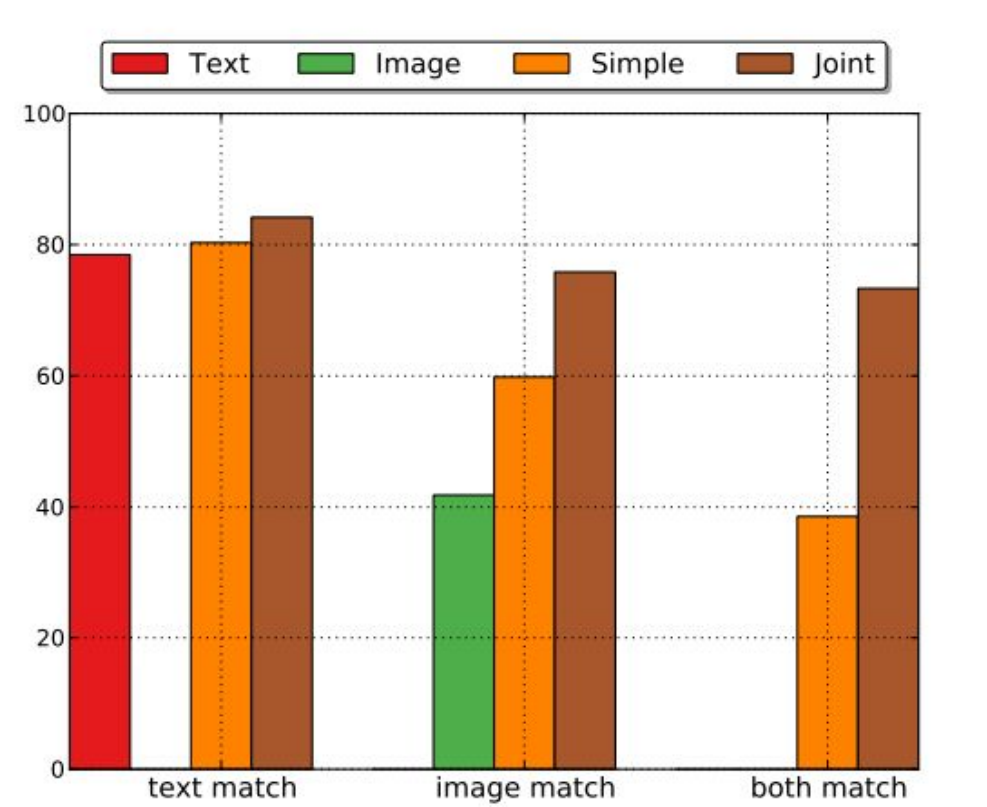
$\tilde{y}^{(2)}$  $\tilde{x}^{(2)}$ 

Simple Text Alignment

- ★ Levenshtein distance
- ★ Fewest number of edits?

Score of 3

1. kitten → sitten (substitution of "s" for "k")
2. sitten → sittin (substitution of "i" for "e")
3. sittin → sitting (insertion of "g" at the end).



We use simple text alignment algorithm.

Process

1. Fetch documents from DDRS database
2. Compute text redaction predictions
3. Treat these predictions as documents
4. Train LDA on predicted redactions
5. Observe word distributions on topics

Dataset

- ★ 213 Document Pairs
- ★ 234 Redaction Predictions

Top 5 Topics, Top 5 Words

0.026*text + 0.025*illegible + 0.017*may + 0.012*military + 0.011*foreign

0.017*british + 0.015*president + 0.014*french + 0.014*council + 0.014*study

0.015*support + 0.015*british + 0.014*committee + 0.014*party + 0.014 *nuclear

0.021*recognized + 0.019*governments + 0.018*united + 0.018*states + 0.015 *military

0.019*war + 0.019*much + 0.018*united + 0.016*shah + 0.015*states

FRUS Documents

Looking at the context of redactions

The Dataset

- ★ Foreign Relations of the United States
- ★ 33K documents for analysis
- ★ Redactions are marked in text:

[1 sentence (2 lines) not declassified]
[1 sentence (2 lines) not declassified]

Contextual Clues

- ★ Extract paragraphs containing redaction markings
- ★ Remove the redaction markings

Documents

- 26,933 docs total
- 923 documents with redactions (3%)
- 1,950 paragraphs in these documents containing redactions
- 4,343 redactions
- Limit by dates 1958-1969 (uneven distribution)

5 Topics (20 Topic LDA)

0.013*meeting + 0.011*soviet + 0.009*targets + 0.009*however + 0.009*possible + 0.009*fy + 0.008*missiles + 0.007*get + 0.007*agency + 0.005*military

0.013*military + 0.011*year + 0.009*government + 0.009*made + 0.009*saigon + 0.007*situation + 0.007*congo + 0.007*policy + 0.007*requirements + 0.007*even

0.010*item + 0.008*present + 0.008*comment + 0.006*support + 0.006*air + 0.006*missile + 0.006*possible + 0.006*war + 0.006*made + 0.006*take

0.019*force + 0.011*program + 0.011*missile + 0.008*immediate + 0.008*matter + 0.008*provide + 0.008*views + 0.007*soviet + 0.007*response + 0.006*missiles

0.014*state + 0.014*saigon + 0.012*force + 0.009*committee + 0.009*request + 0.009*project + 0.009*party + 0.007*situation + 0.007*department + 0.007*risk

FRUS Documents:

Looking for Redacted Topics

What are “Hot” Topics?

1. Run LDA on full documents (40 topics)
2. Assign ea. doc to most likely topic
3. Find number of redactions in each doc
4. FInd topics with most redactions.

Sample Topics across All

1. 0.070*israel + 0.033*arab + 0.030*israeli + 0.020*jordan + 0.018*israelis + 0.018*arabs + 0.012*middle + 0.011*east + 0.009*arms + 0.006*water
2. 0.146*president + 0.016*nixon + 0.013*presidentnixon + 0.013*kissinger + 0.013*asked + 0.010*vice + 0.009*discussion + 0.008*meeting + 0.006*dr + 0.005*memorandum
3. 0.027*meeting + 0.010*state + 0.007*morning + 0.007*time + 0.006*telegram + 0.006*statement + 0.005*president + 0.005*message + 0.005*meetings + 0.005*made
4. 0.022*united + 0.016*resolution + 0.014*states + 0.010*assembly + 0.009*nations + 0.007*new + 0.007*general + 0.007*position + 0.006*council + 0.006*treaty
5. 0.017*agreement + 0.011*proposal + 0.011*negotiations + 0.009*position + 0.007*talks + 0.006*agreed + an + 0.006*state + 0.006*rebel + 0.005*support + 0.005*central + 0.005*situation + 0.005*army

Results

Most heavily redacted topic has 535 redations

'0.014*political + 0.011*military + 0.008*communist + 0.007*probably + 0.007*government + 0.006*support + 0.006*may

+ 0.005*power + 0.005*economic + 0.005*policy'

Resources

Topic Modelling code
iPython Notebook examples
declass-align/topicmodel
(see results dir)
“Topic Modelling README”

Credits

Topic Modelling:

- ★ https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf
- ★ http://mcburton.net/blog/joy-of-tm/