Text Mining & Search

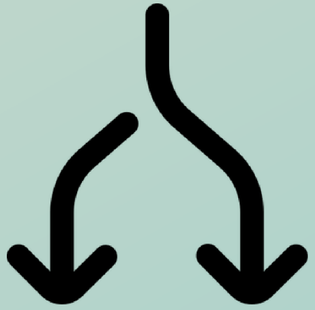
Alberto Gadda **824029**

Paolo Guerini Rocco **826236**

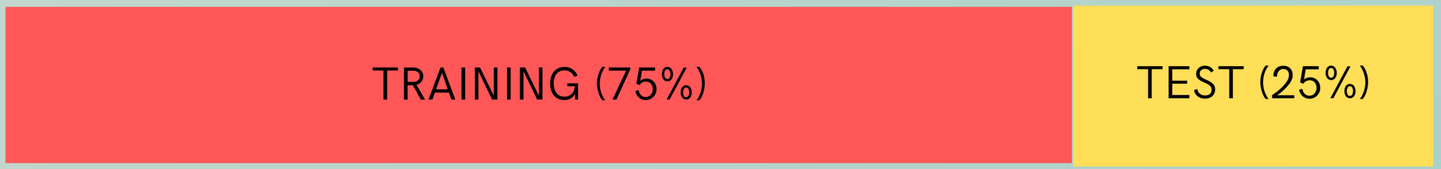
Data from three different sources

	Number of emails	Percentage of spam
• Assassin dataset	6046	31 %
• Enron dataset	10000	50 %
• Ling dataset	2605	17 %
• Total	18651	

Training-Test split



- Training: 13988 email
- Test: 4663 email



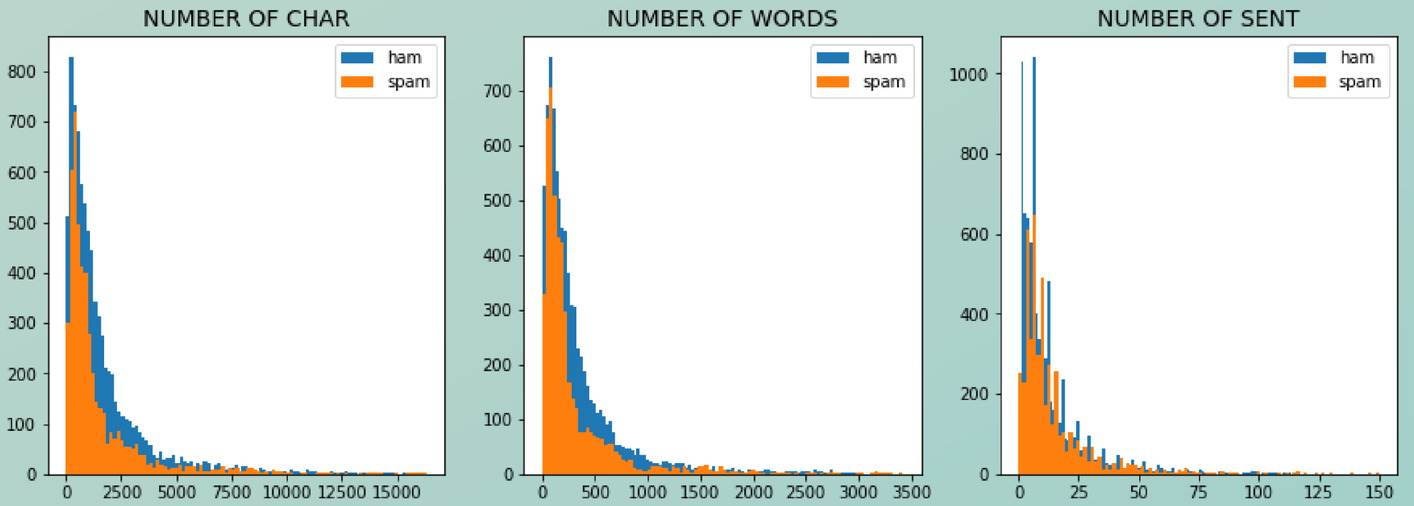
Partitioning made by stratified random sampling based on labels

Metrics chosen for performance evaluation:

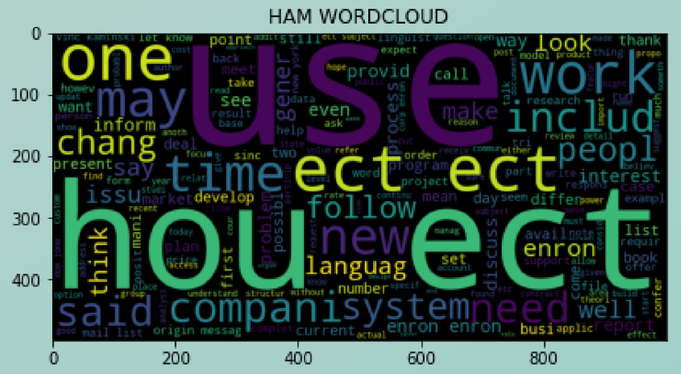
1. Accuracy (dataset is not heavily imbalanced)
2. Precision (false positives are detrimental)
3. Recall

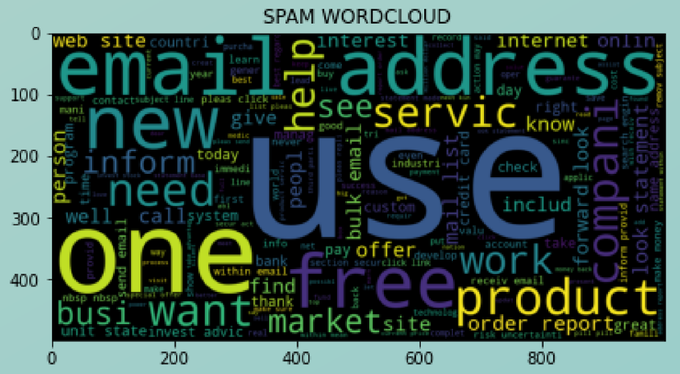
Exploratory Analysis

- Distribution of document length

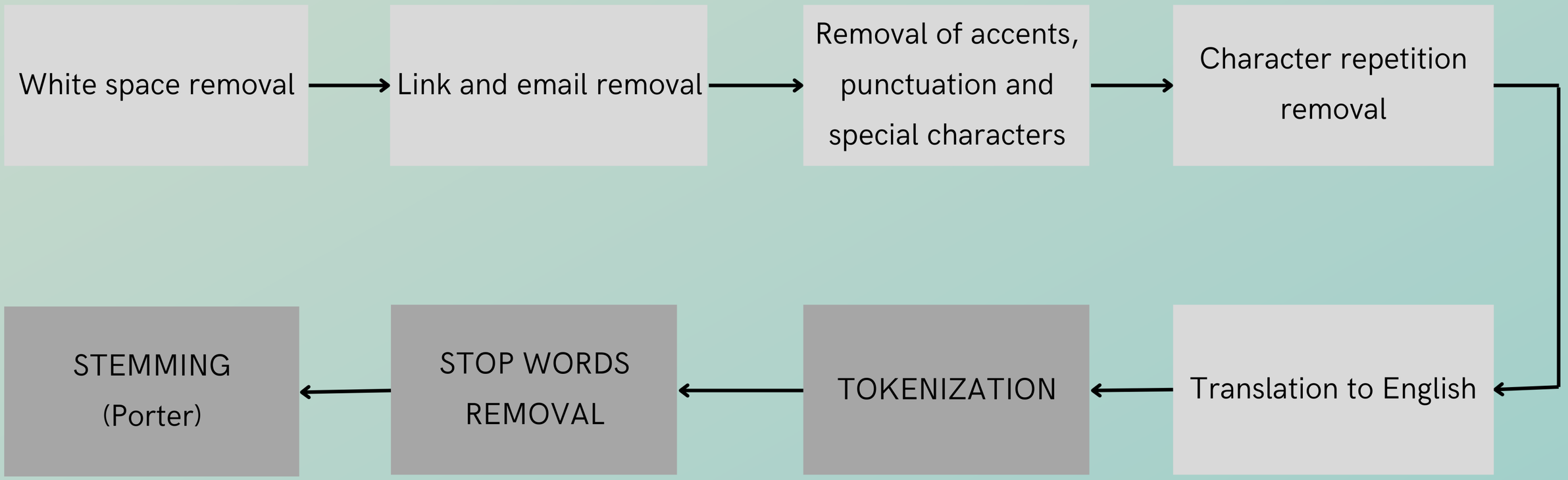


- Wordcloud of more frequent terms





Preprocessing



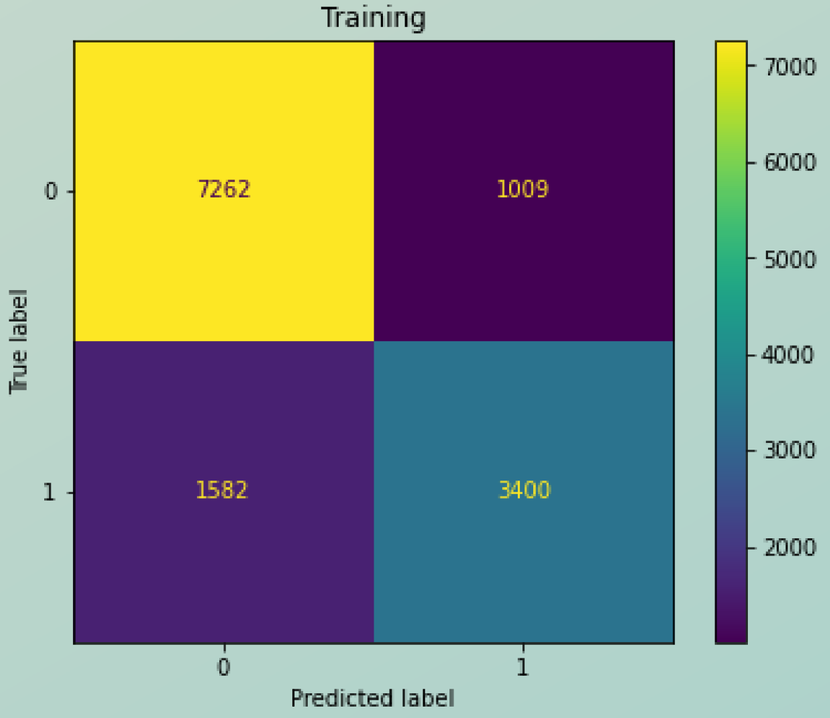
CLASSIFICATION

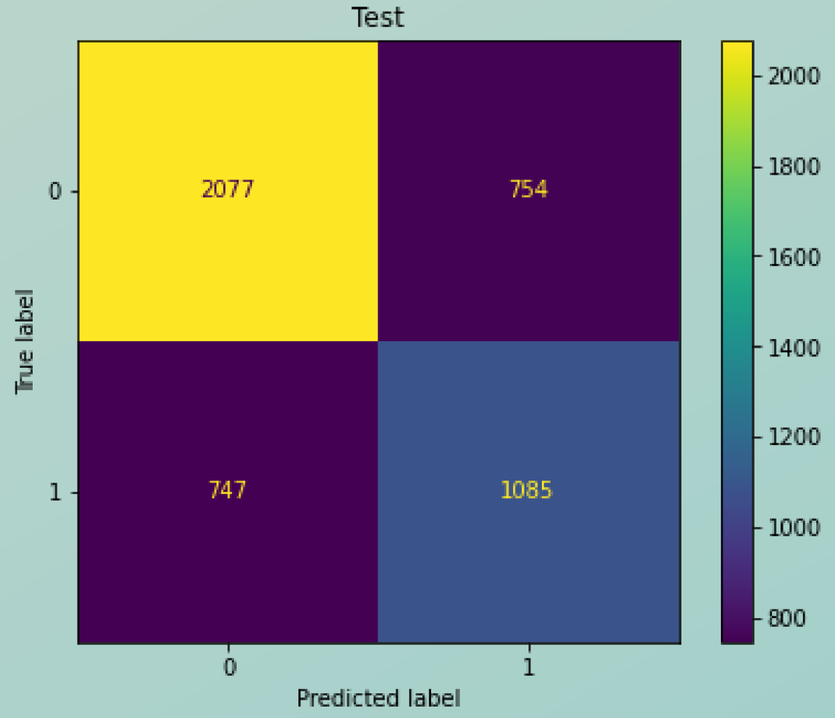
Bag-of-Words Document Representation:

- TF-IDF vectorizer (minimum reqired term frequency: 30)
- PCA (first 500 components)
- Additional features:
  - # of words
  - # of characters
  - # of sentences
  - # of special characters removed
  - If any links have been removed from the original text
  - If any email addresses have been removed from the original text
  - Levenstein distance between before and after preprocessing

Naive Bayes

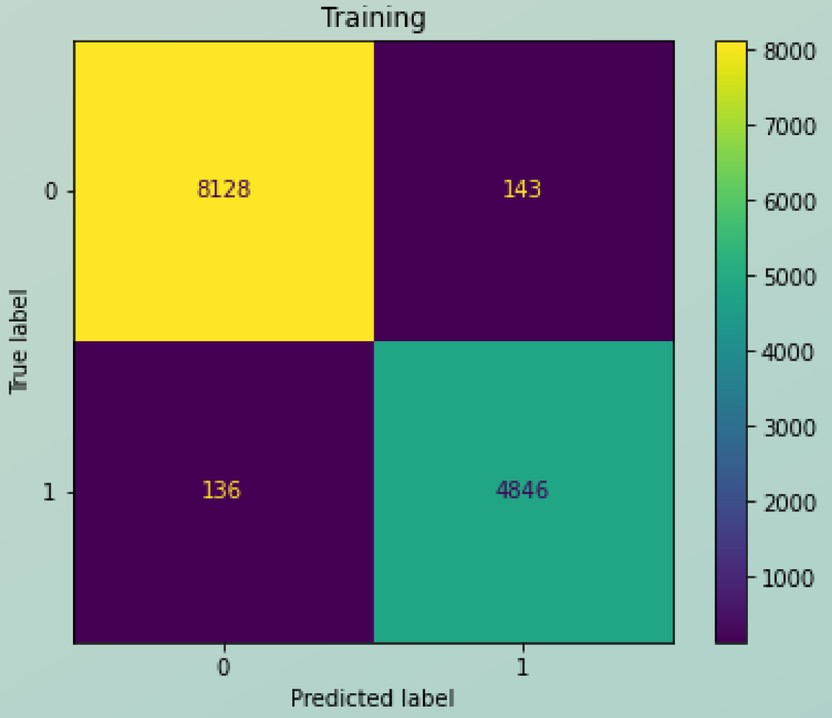
	ACCURACY	PRECISION	RECALL
TRAIN	0.8	0.77	0.68
TEST	0.68	0.59	0.59

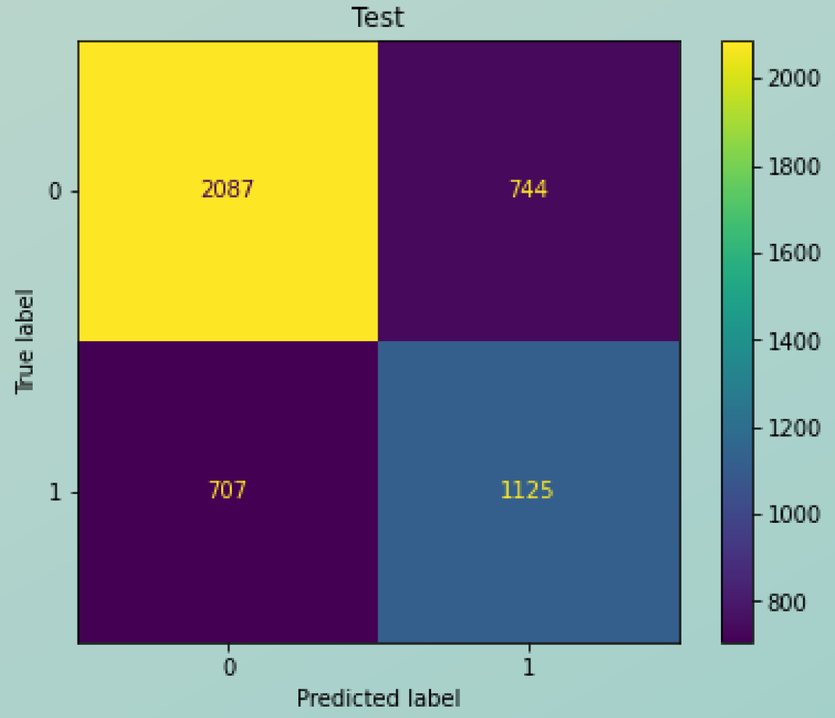




XGBoost

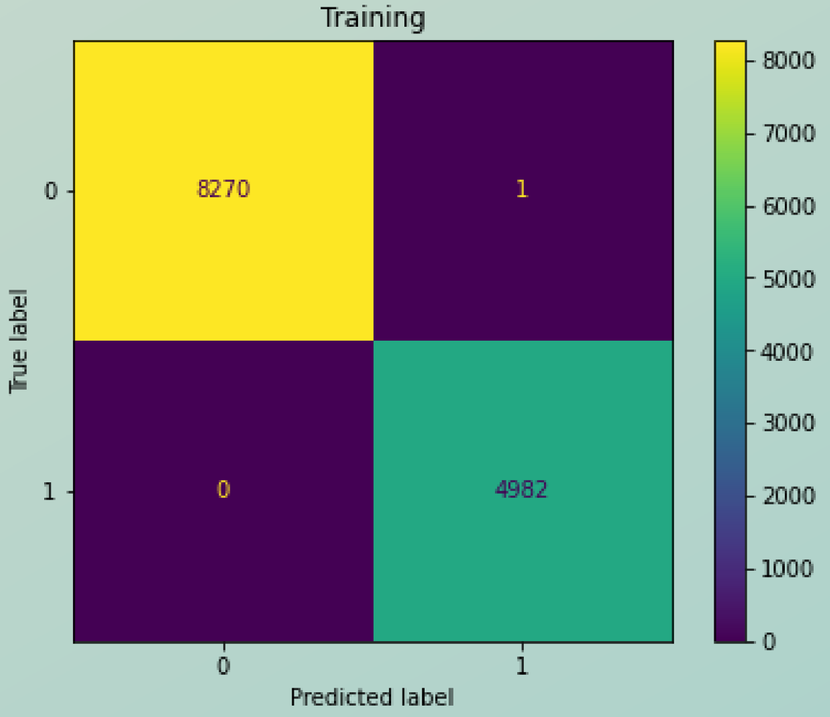
	ACCURACY	PRECISION	RECALL
TRAIN	0.98	0.97	0.97
TEST	0.69	0.60	0.61

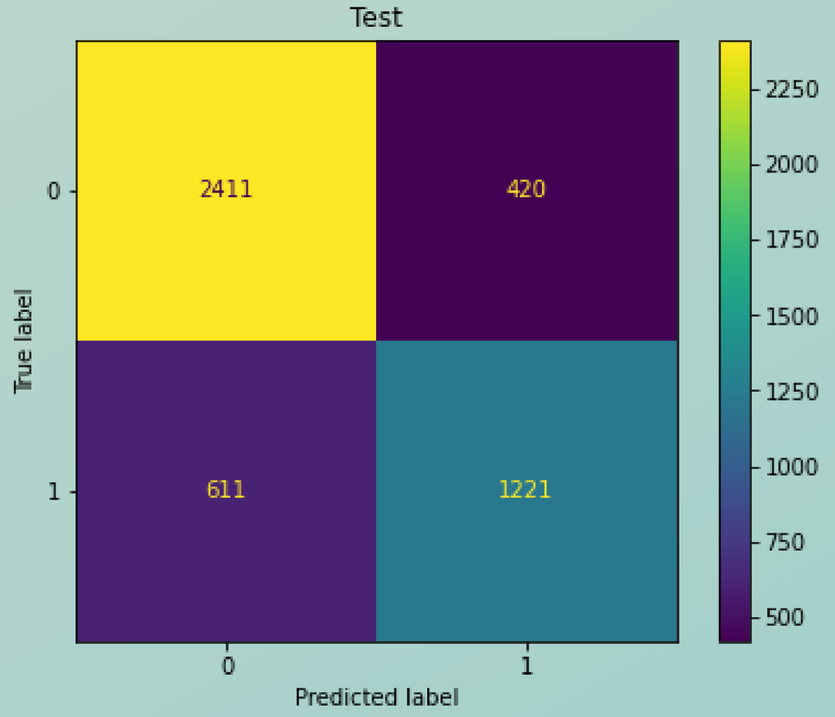




MLP

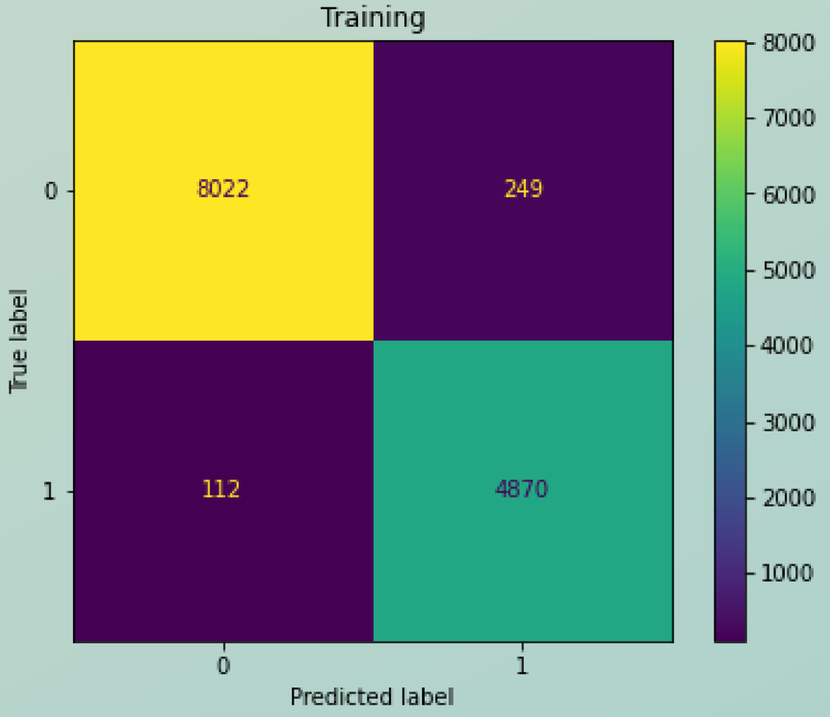
	ACCURACY	PRECISION	RECALL
TRAIN	1.00	1.00	1.00
TEST	0.78	0.74	0.67

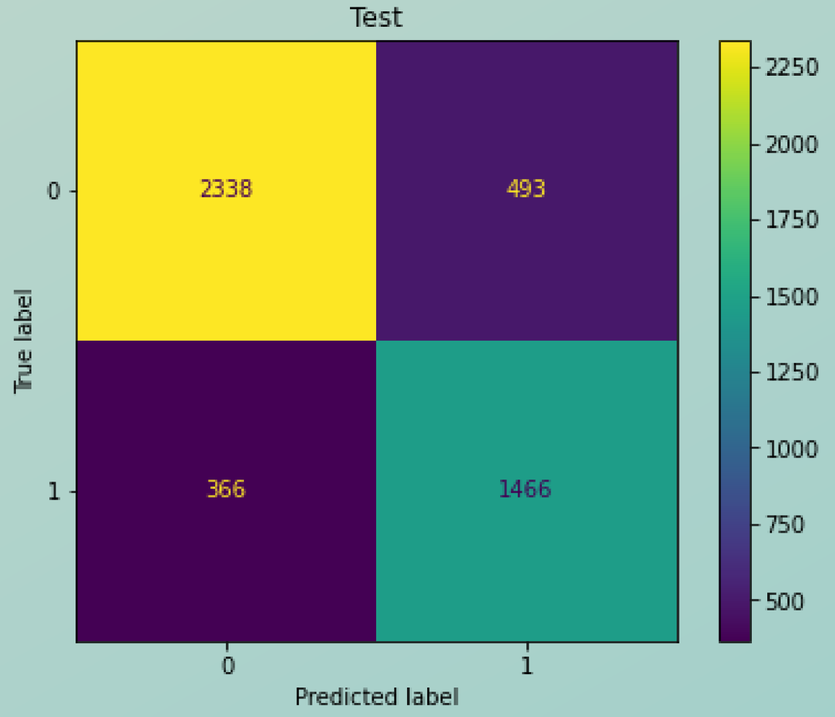




SVC

	ACCURACY	PRECISION	RECALL
TRAIN	0.97	0.95	0.98
TEST	0.82	0.75	0.80





TOPIC MODELING

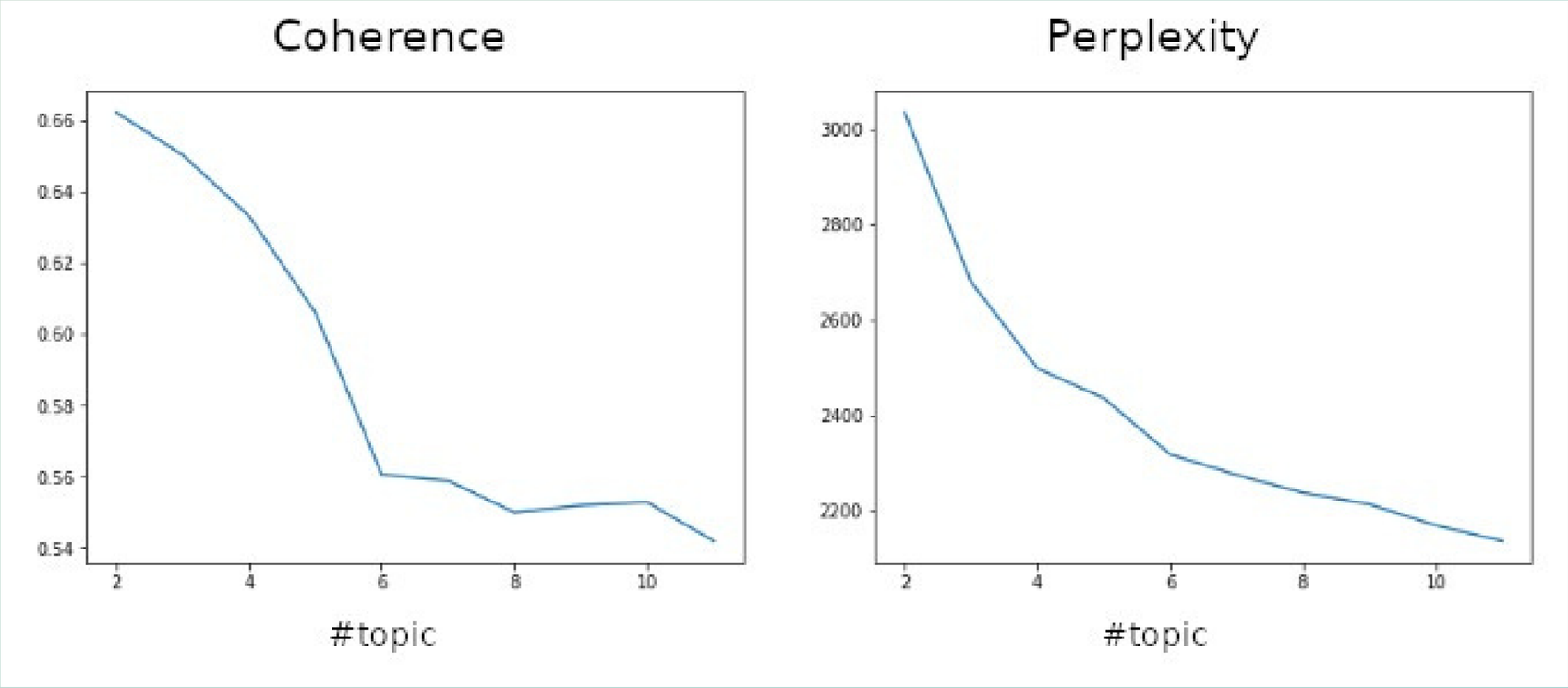
DOCUMENT REPRESENTATION:

- Count vectorizer
  - Minimum required term frequency: 5
  - Maximum term frequency: presence in 50% of documents

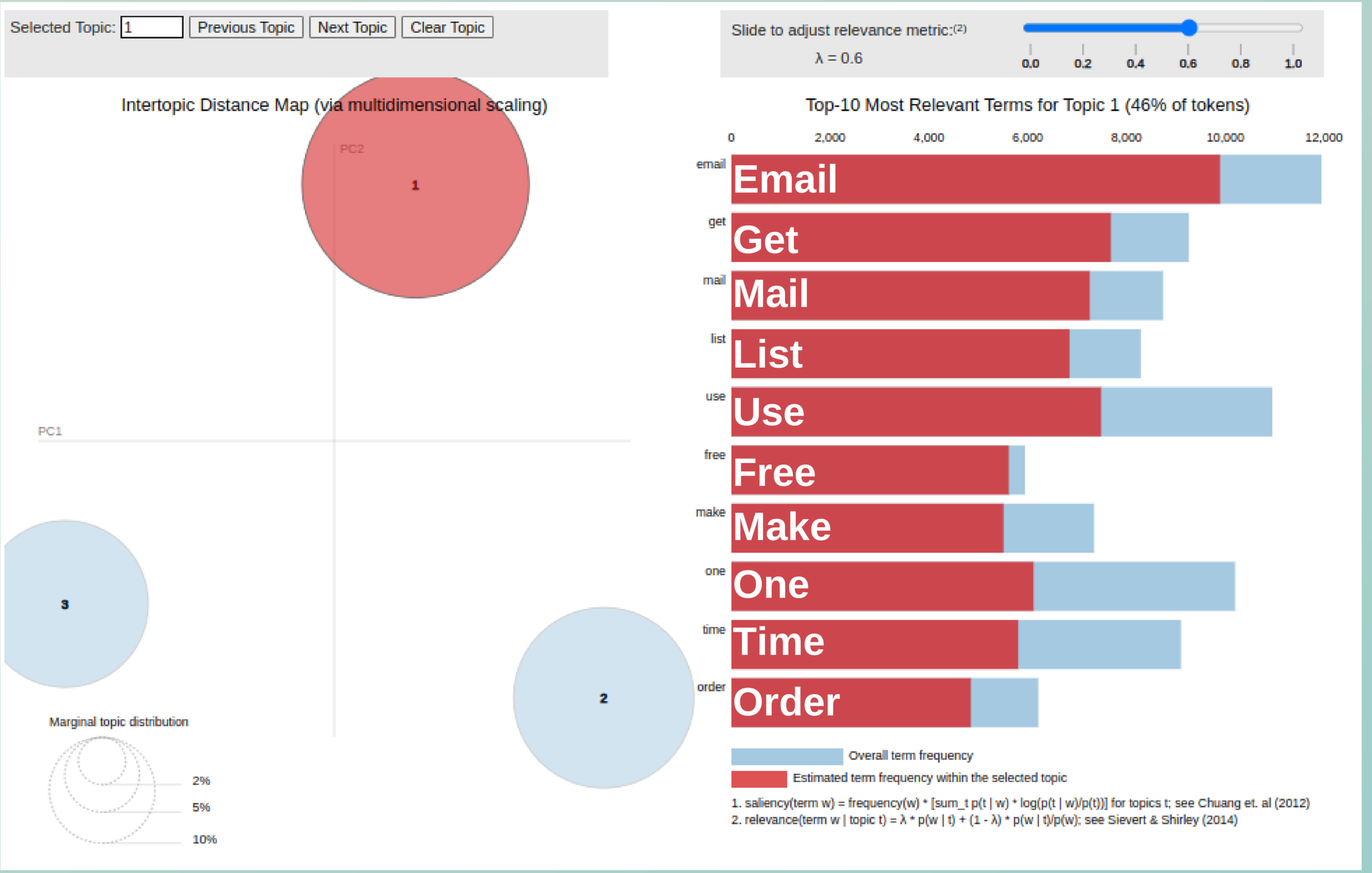
- Documents are represented by a vector of 21978 elements

Topic Numerosity Selection

- Coherence
- Perplexity



Topic #1 (λ=0.6)


Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.6
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
2%
5%
10%
Top-10 Most Relevant Terms for Topic 1 (46% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
email
Email
get
Get
mail
Mail
list
List
use
Use
free
Free
make
Make
one
One
time
Time
order
Order
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic #2 (λ=0.6)

Selected Topic: 2 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

λ = 0.6

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)

PC2

PC1

1

2

3

Marginal topic distribution

2%

5%

10%

Top-10 Most Relevant Terms for Topic 2 (29.1% of tokens)

0 5,000 10,000 15,000 20,000

enron Enron

ect Ect

compani Compani

hou Hou

said Said

energi Energy

trade Trade

subject Subject

deal Deal

market Market

Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Topic #3 (λ=0.6)

Selected Topic: 3 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2)

λ = 0.6

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)

PC2

PC1

1

2

3

Marginal topic distribution

2%

5%

10%

Top-10 Most Relevant Terms for Topic 3 (24.8% of tokens)

0 2,000 4,000 6,000 8,000

languag Languag

univers Univers

linguist Linguist

paper Paper

confer Confer

edu Edu

english English

workshop Workshop

research Research

studi Studi

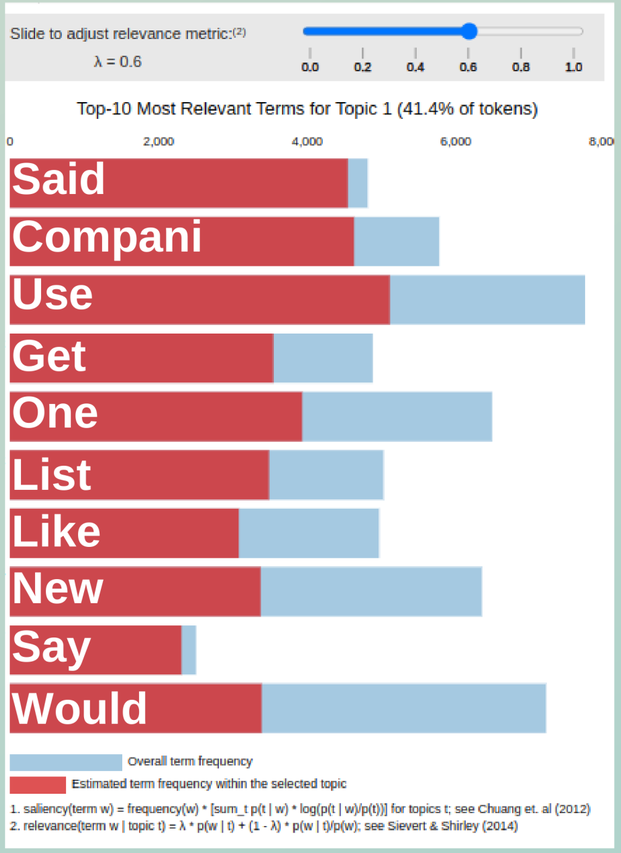
Overall term frequency

Estimated term frequency within the selected topic

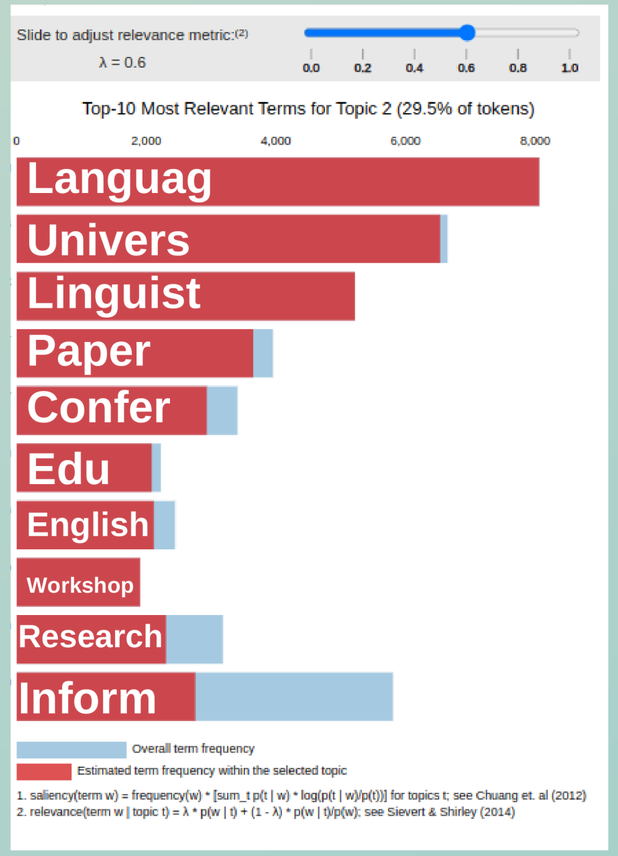
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Only Ham (λ=0.6)

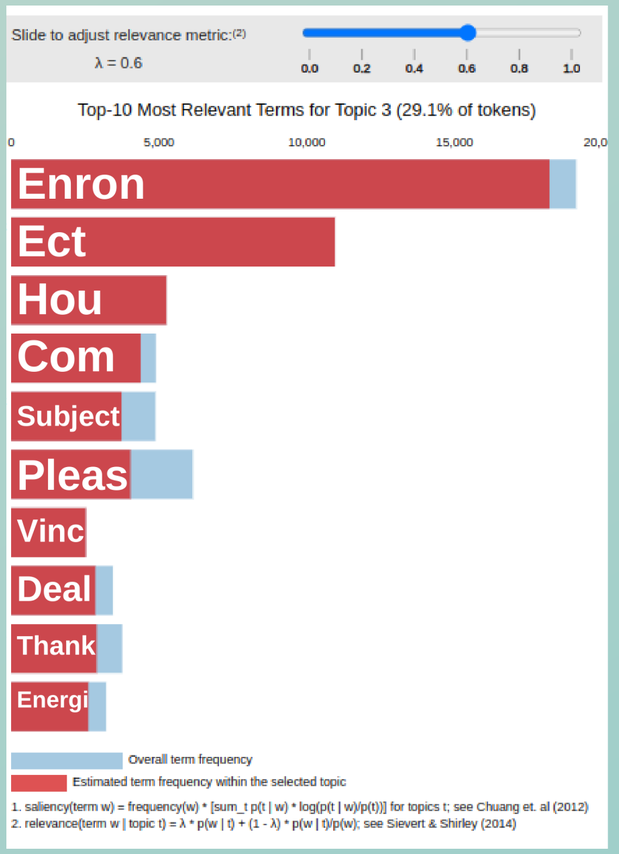
Topic 1



Topic 2



Topic 3

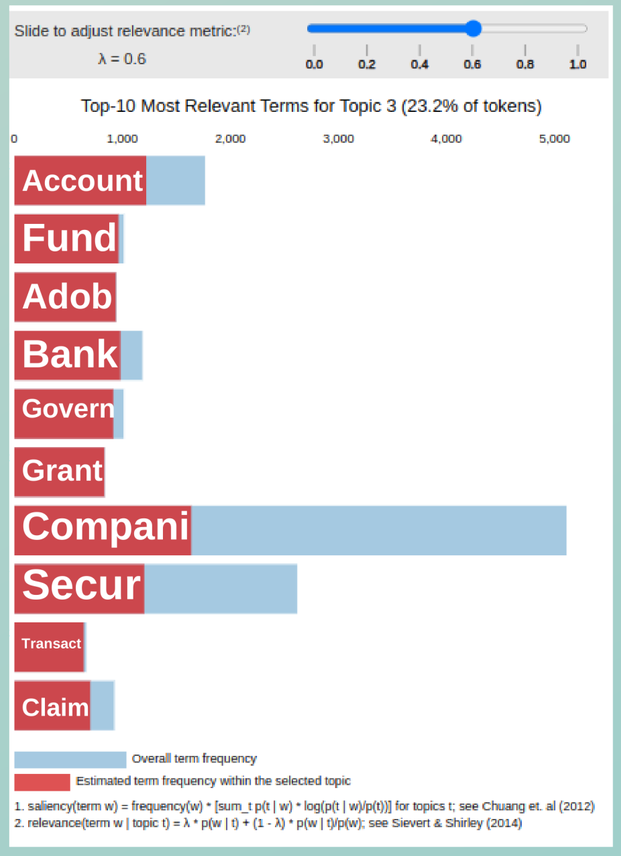


Only Spam (λ=0.6)

Topic 1

Topic 2

Topic 3



Topic-Source Relationship

Are the identified topics attributable to any single dataset component?

HAM	Dataset	% of Documents from this dataset
TOPIC 1	Assassin	90
TOPIC 2	Ling	85
TOPIC 3	Enron	99

SPAM	Dataset	% of Documents from this dataset
TOPIC 1	Enron	67
TOPIC 2	Assassin	25
TOPIC 3	Enron	88


THANK
YOU!