What’s in the mail box?

Filippo Medri
Metis 2018 San Francisco

AGENDA

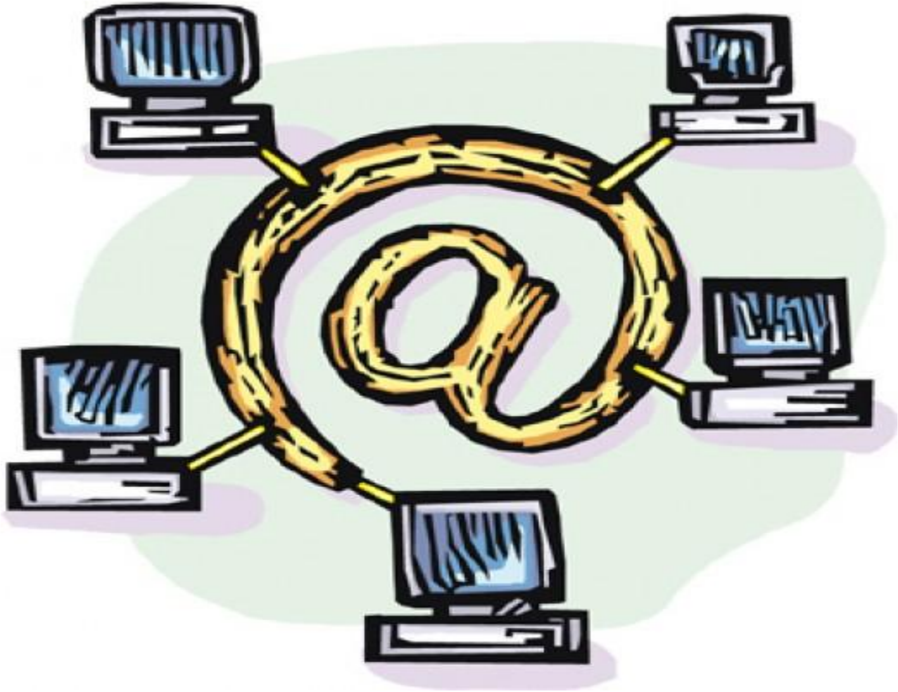
❖ 3000 e-mails

➢ LSA + K-Means
➢ LDA

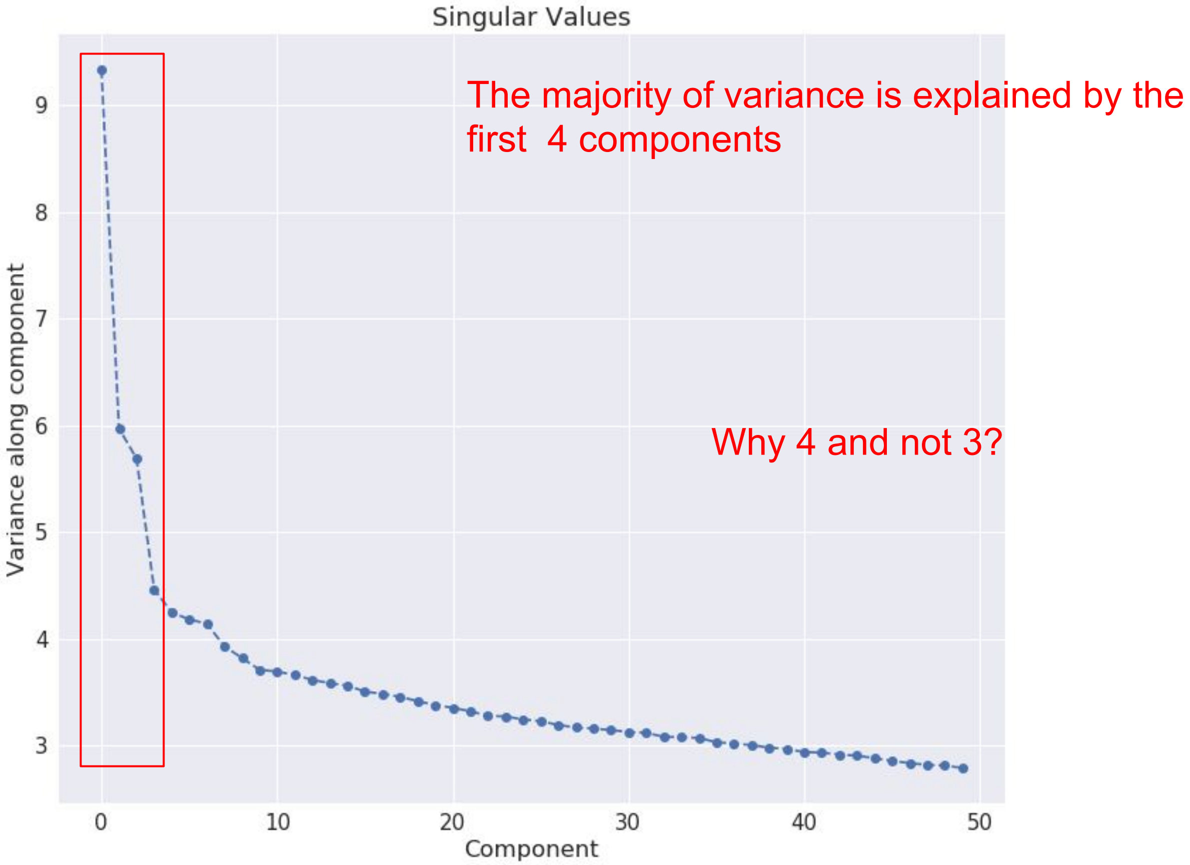
➔ Topic Discovery
➔ Automatic Categorization

20Newsgroups Dataset

- **3** newsgroups :
  - comp.graphics
  - rec.autos
  - talk.misc.religion
- Removed **header, footers, quotes**

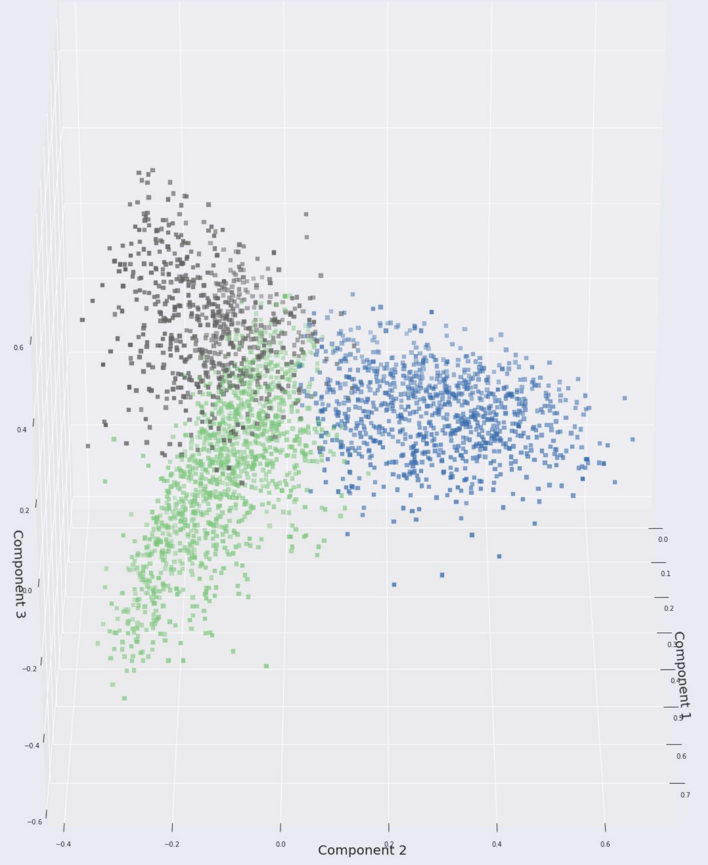


LSA - 50 components from 1000 features



3-Means

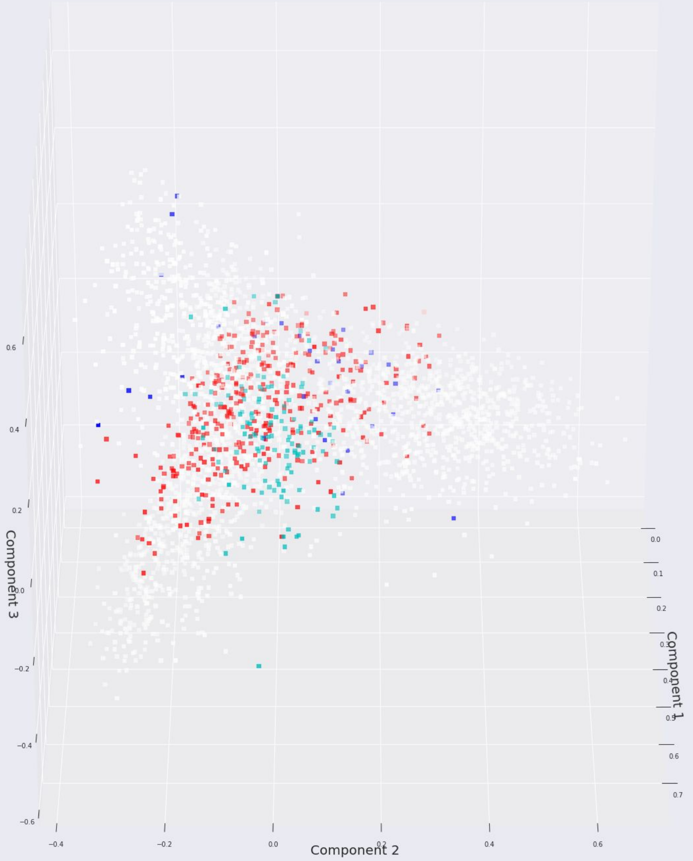
PEOPLE	thank	car
GOD	graphic	engine
THINK	image	new
CHRISTIAN	know	drive
SAY	program	good
KNOW	computer	speed
LIFE	looking	ford
MAKE	please	dealer
GOOD	file	problem
JESUS	use	price



Misclassification

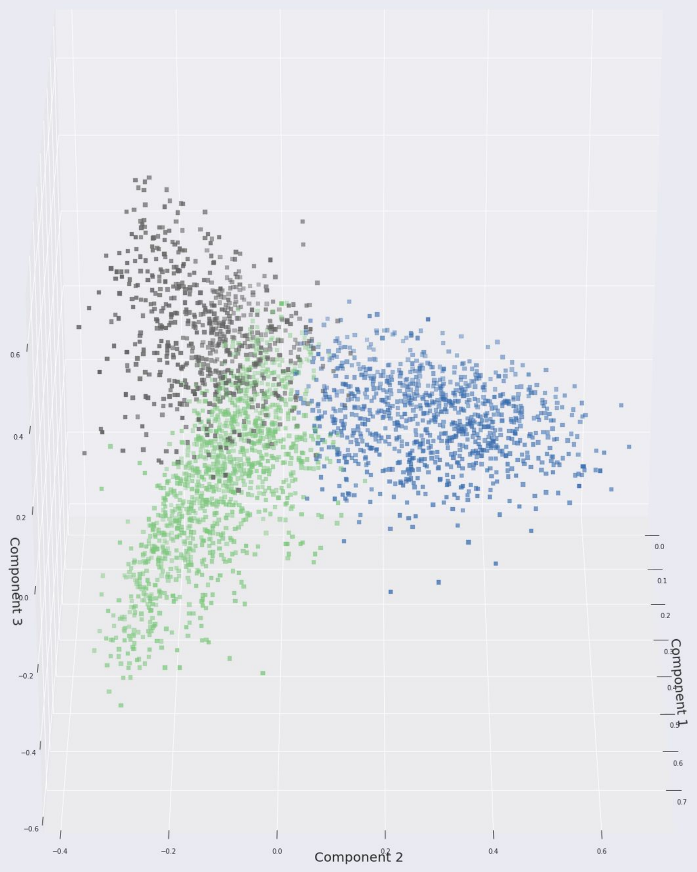
Overall	**17 %**
rec.autos	*35.3 %*
comp.graphics	12.6 %
TALK.MISC.RELIGION	3.4 %

WHITE: correctly classified

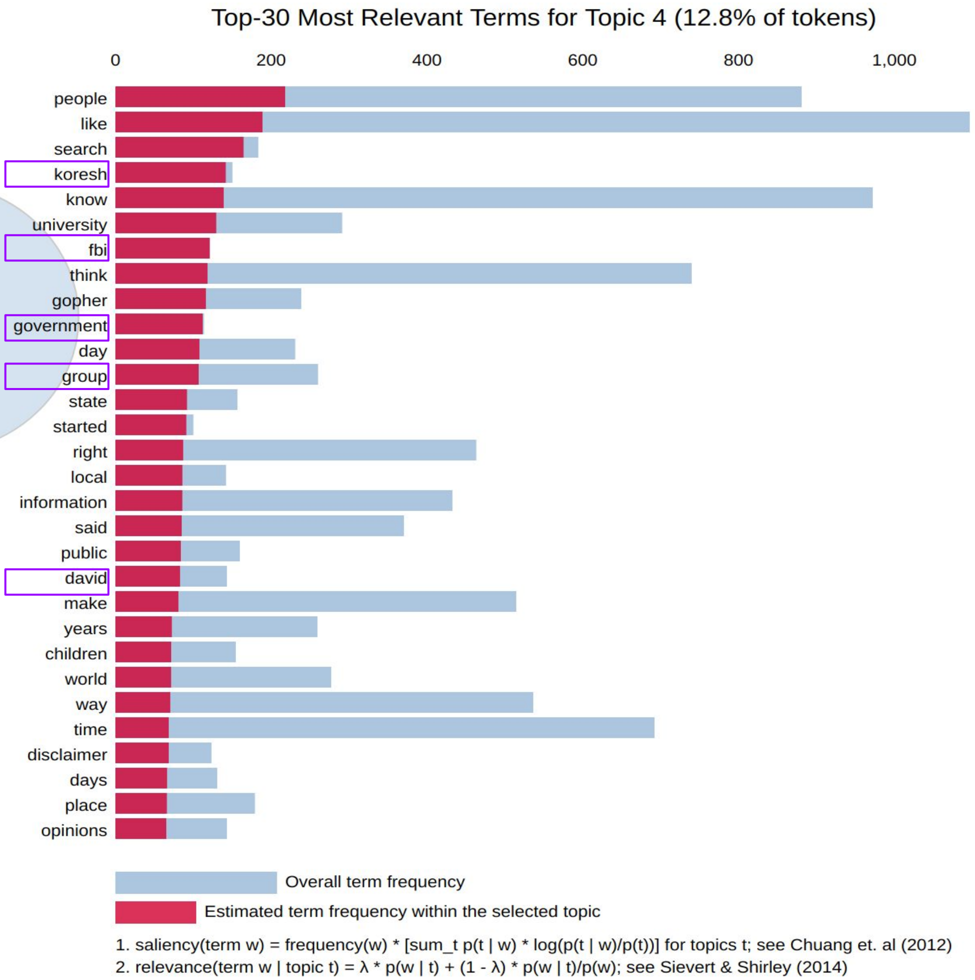


3-Means

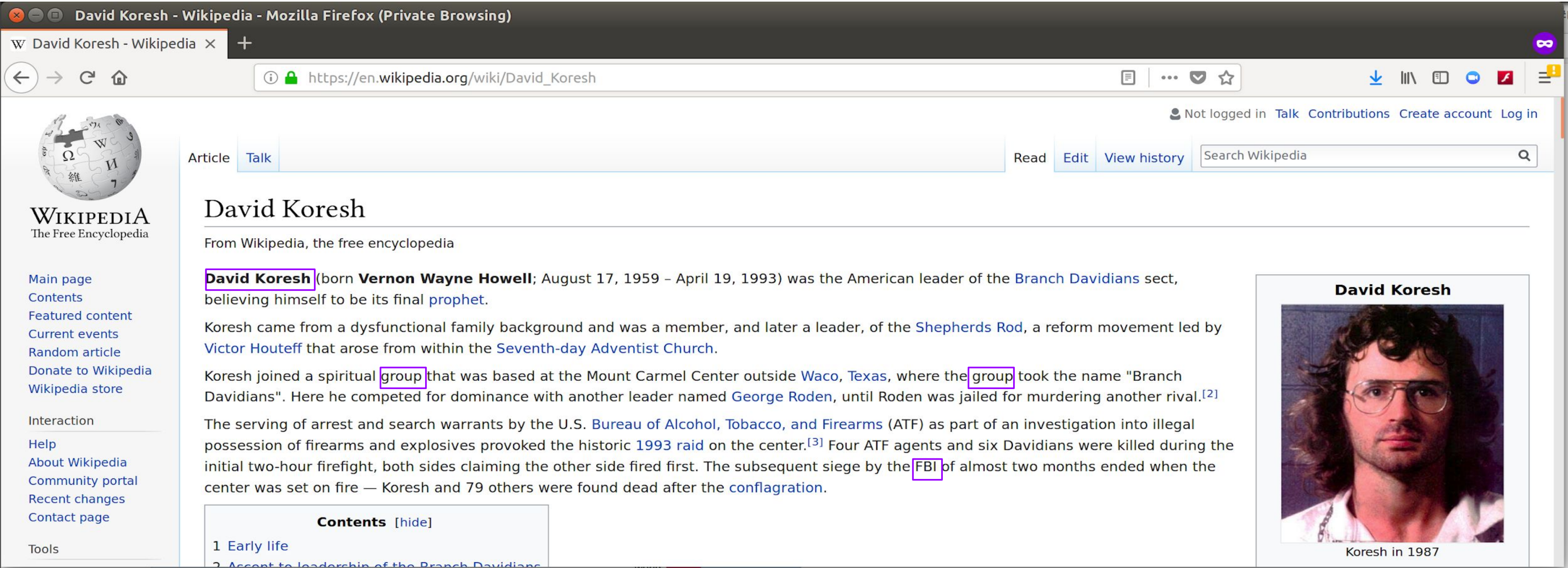
PEOPLE	thank	car
GOD	graphic	engine
THINK	image	new
CHRISTIAN	know	drive
SAY	program	good
KNOW	computer	speed
LIFE	looking	ford
MAKE	please	dealer
GOOD	file	problem
JESUS	use	price

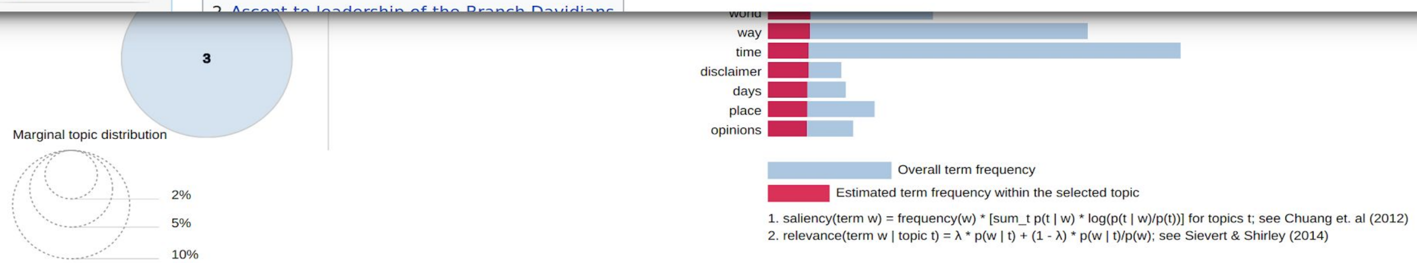


LDA - Unaspected topic ….



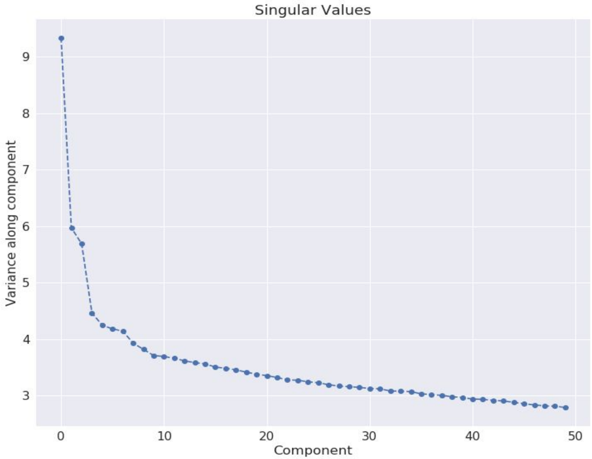
LDA - it is about David Koresh !!!





Conclusions

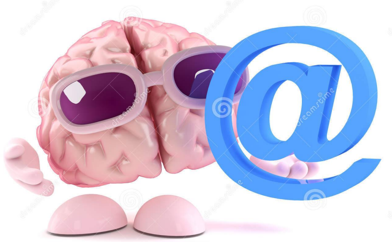
Component distribution



Automatic categorization

Discovery of **new** topics !!!

Pre-training of deep learning models

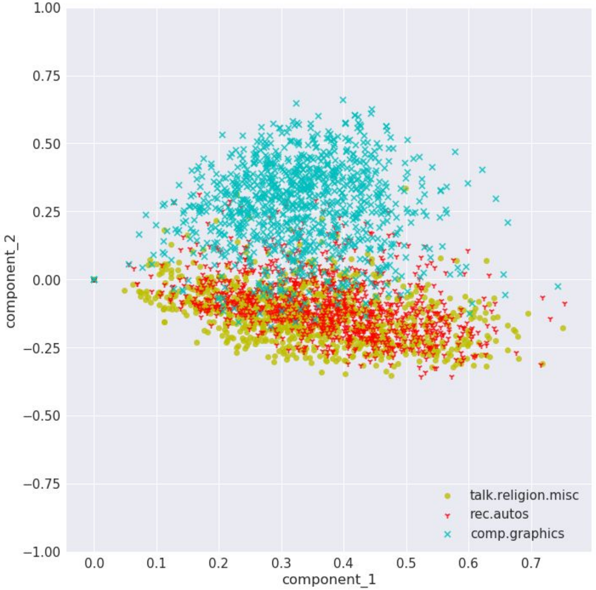


Thank You!!

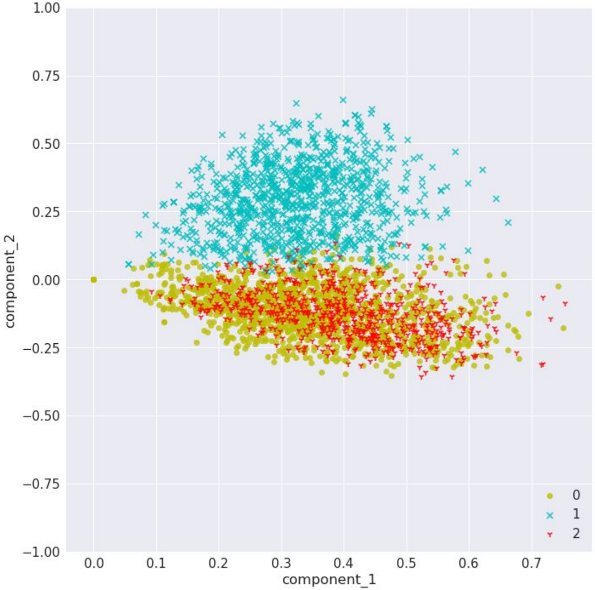
Filippo Medri
Metis 2018 San Francisco

Discard Pile

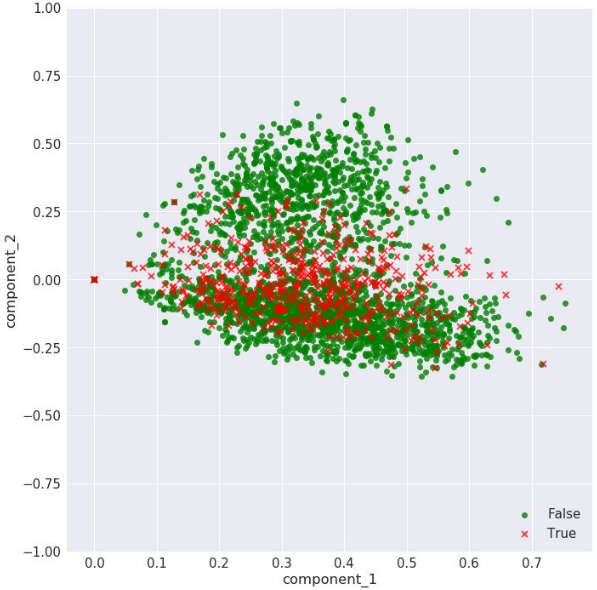
Clusters vs Category on first 2 principal components



Original Categories

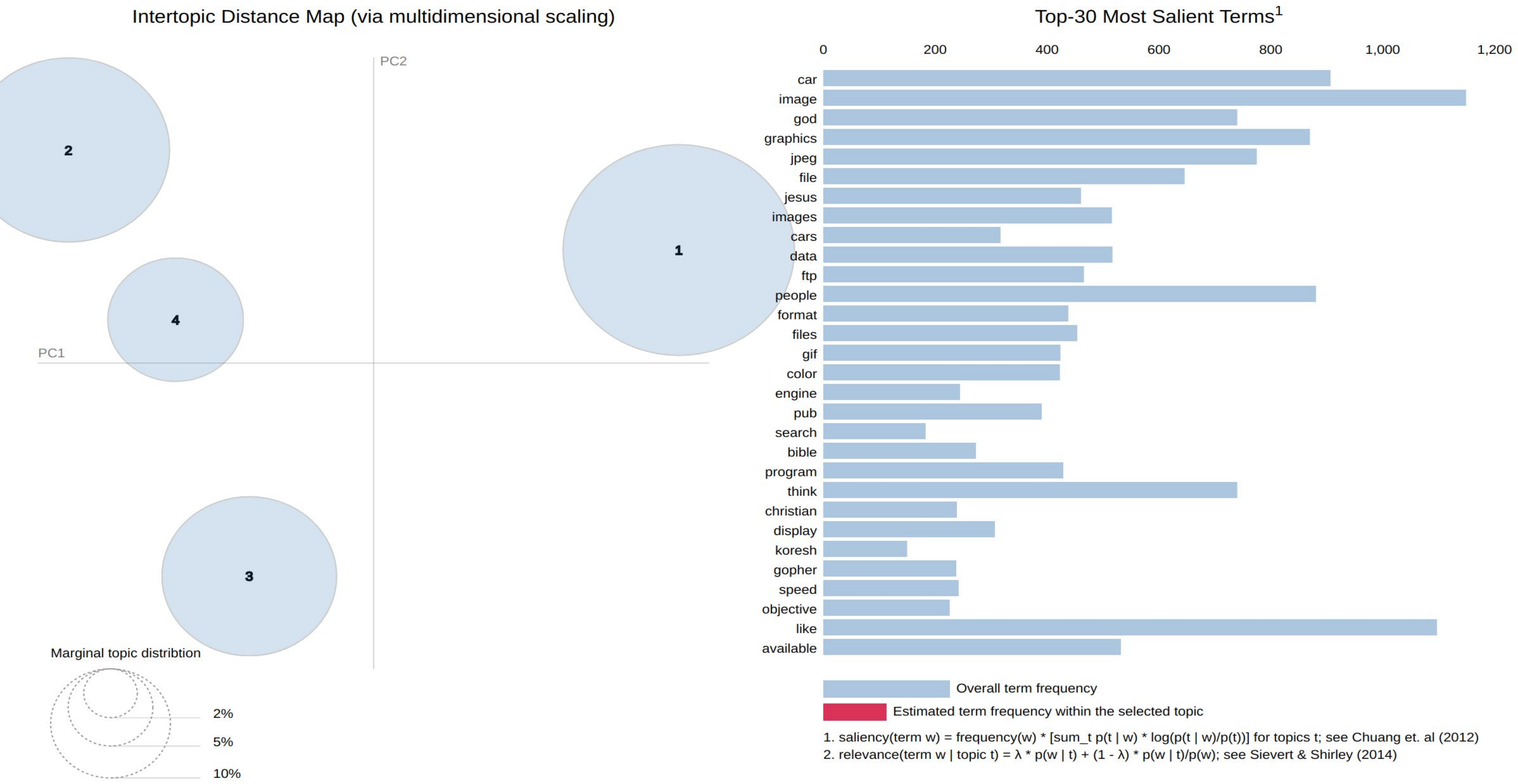


Clusters

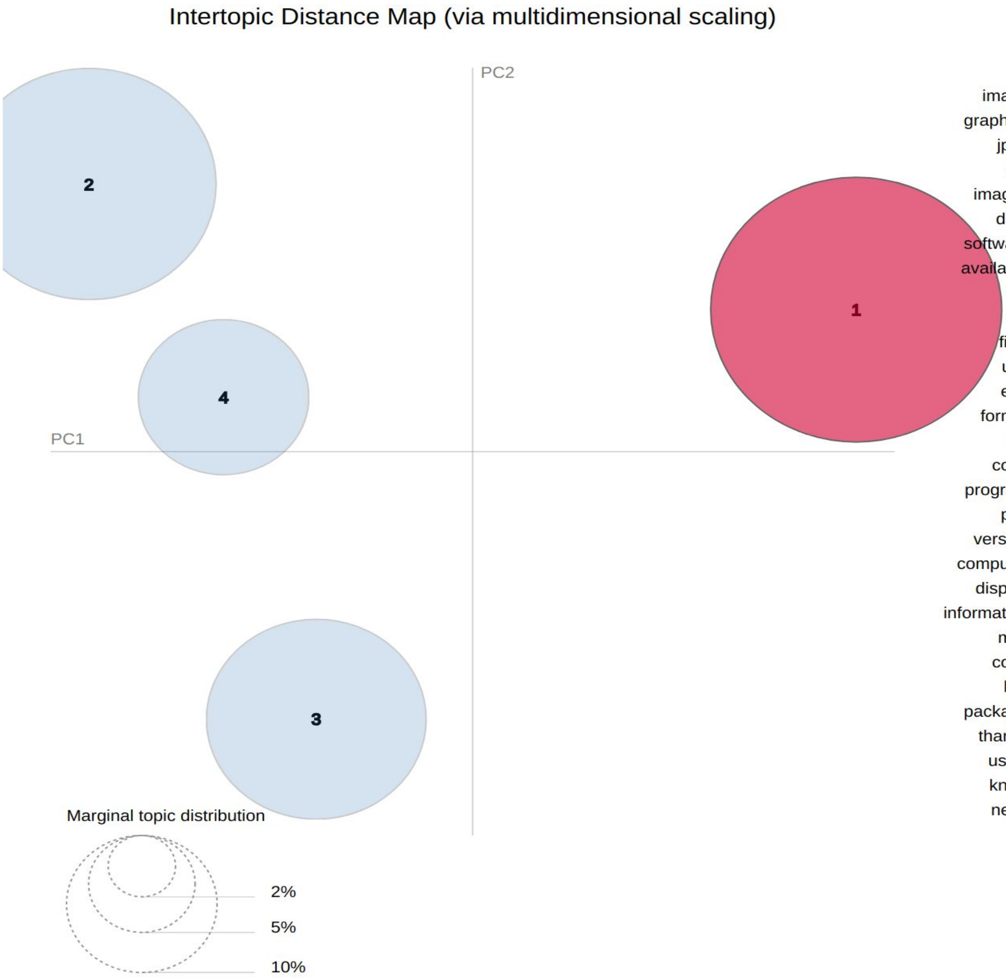


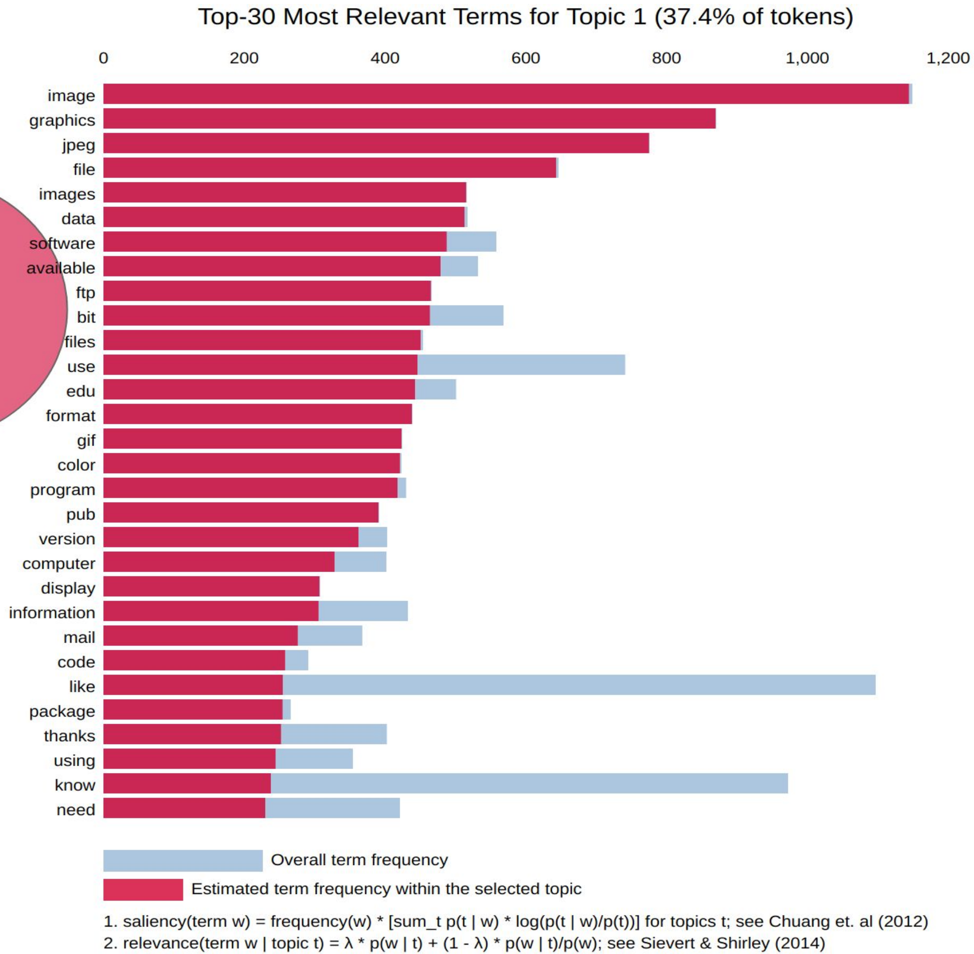
Misclassified

LDA - Overall

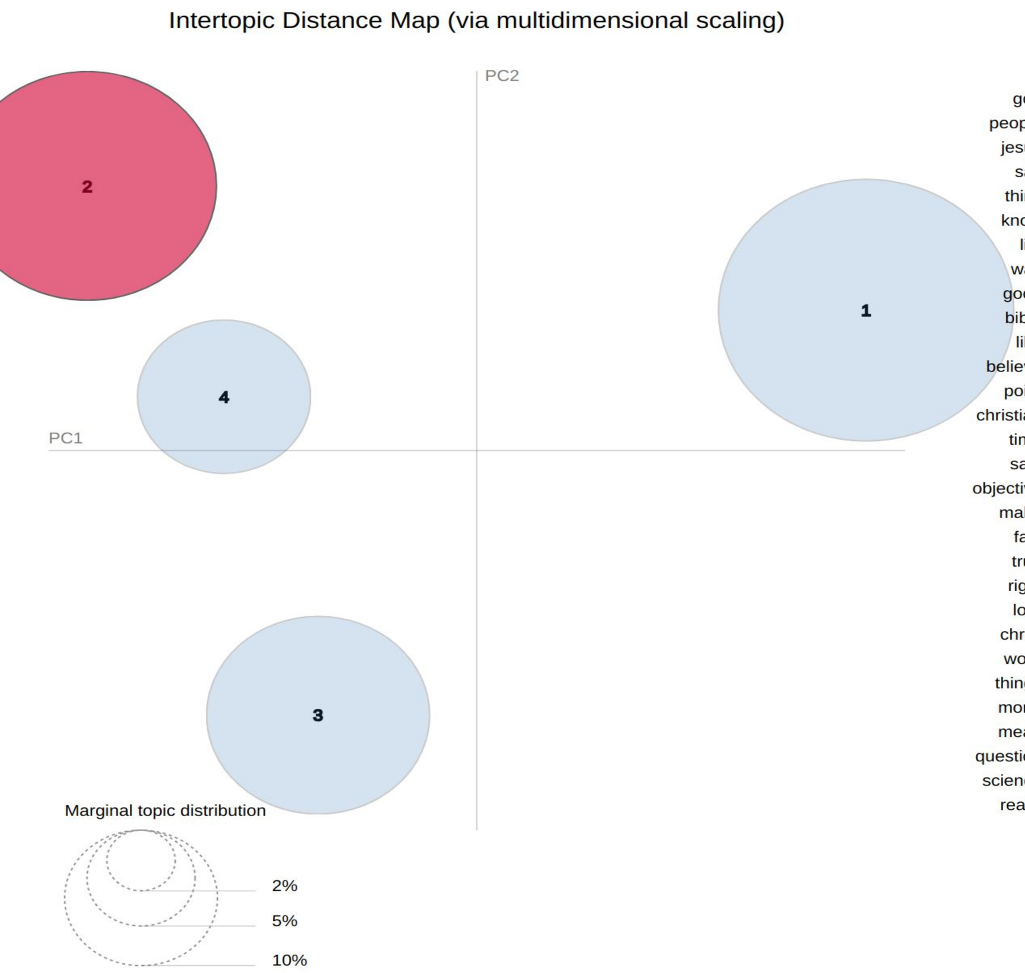

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribtion
2%
5%
10%
Top-30 Most Salient Terms1
0
200
400
600
800
1,000
1,200
car
image
god
graphics
jpeg
file
jesus
images
cars
data
ftp
people
format
files
gif
color
engine
pub
search
bible
program
think
christian
display
koresh
gopher
speed
objective
like
available
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)

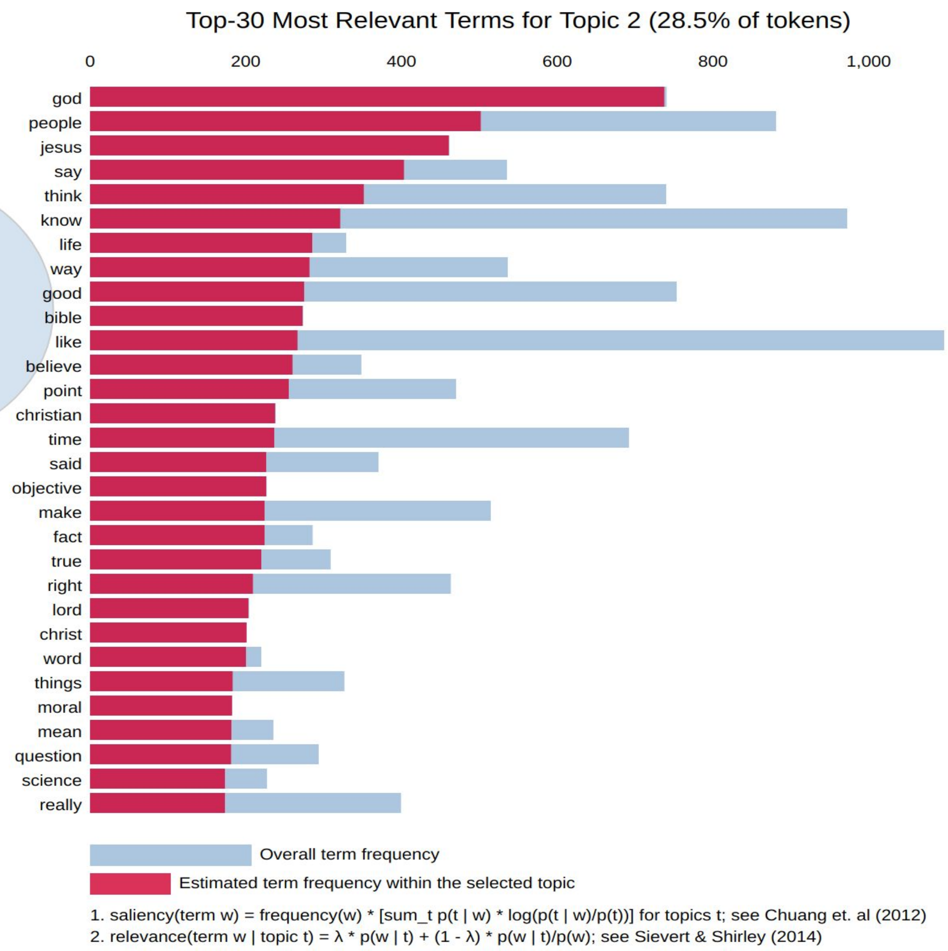
LDA - Graphics



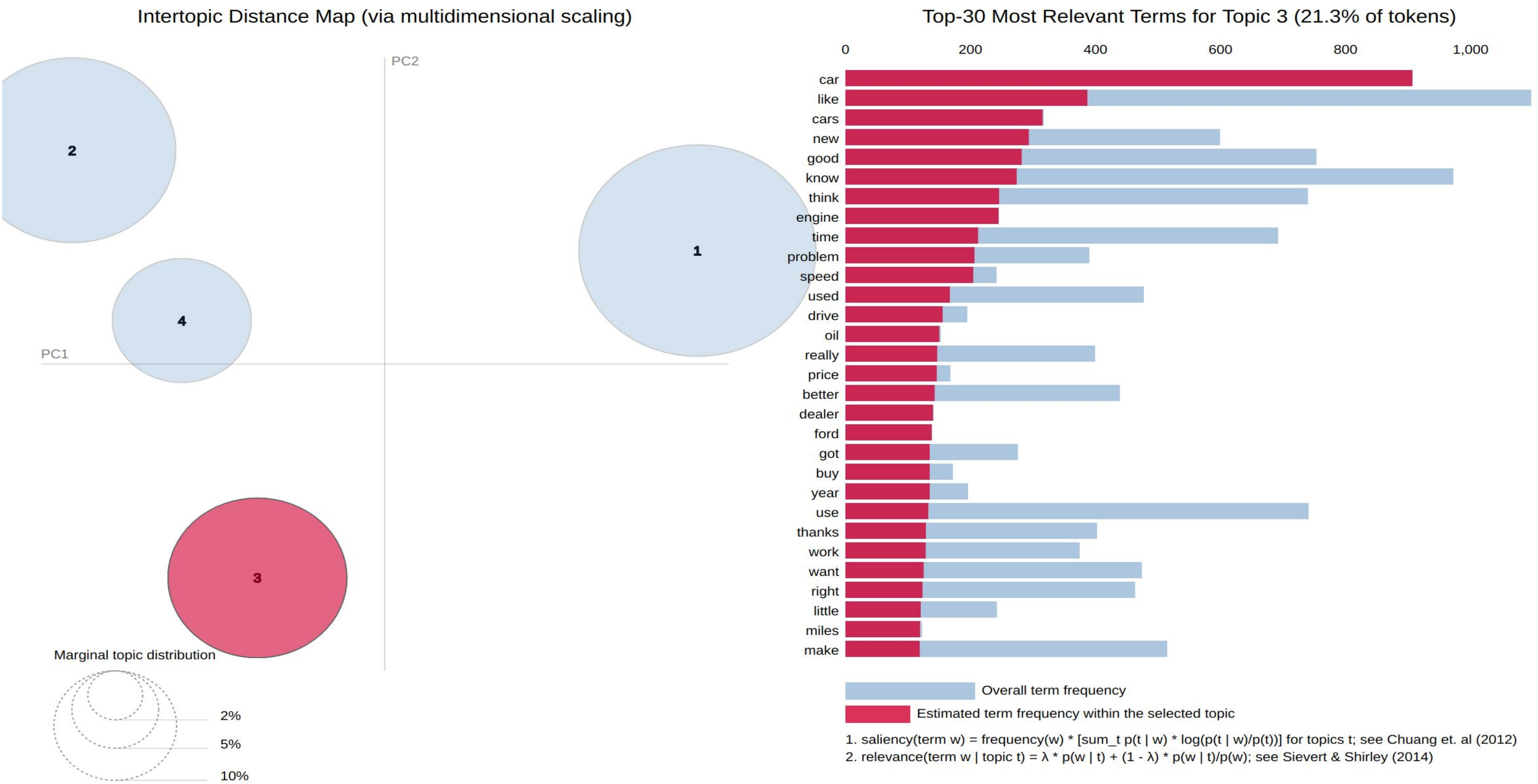


LDA - Religion





LDA - CAR


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
4
1
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 3 (21.3% of tokens)
0
200
400
600
800
1,000
car
like
cars
new
good
know
think
engine
time
problem
speed
used
drive
oil
really
price
better
dealer
ford
got
buy
year
use
thanks
work
want
right
little
miles
make
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)

*Can we extract the mail **topics** from a mail box?*

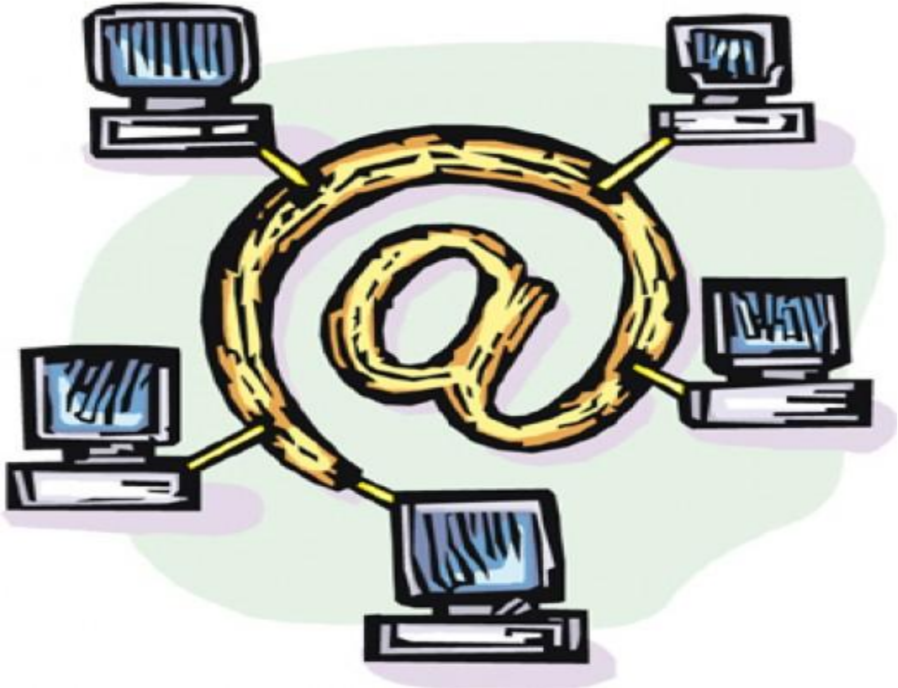
Let’s try with LSA+K-means and LDA !

Can we extract the topics from a mail box?

- **Label each e-mail to one topic**
- Discover hidden topics

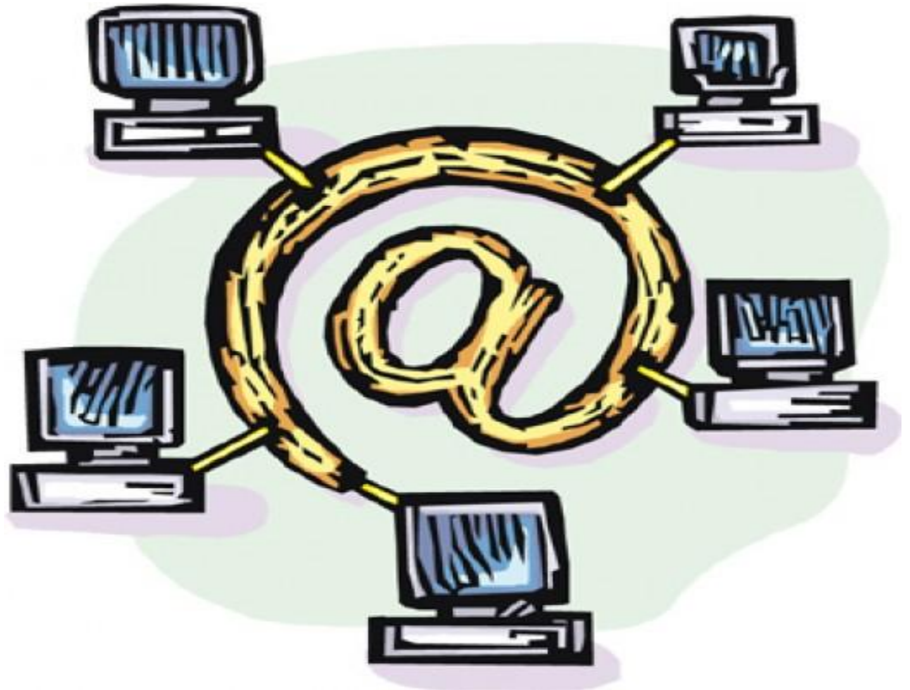
- 20NewsGroups Dataset

- LSA + K-Means
- Latent Dirichlet Analysis



Can we extract the topics from a mail box?

- **3000** e-mails from the 20Newsgroups dataset
- **3 newsgroups** :
  - comp.graphics
  - rec.cars
  - talk.misc.religion
- **2 modeling attempts:**
  - Dimensionality reduction through LSA + Clustering through K-Means
  - Latent Dirichlet Analysis
- **1.5 Goal(s) :**
  - **Automatic categorization of each e-mail to one of the 3 newsgroups main topic**
  - Find the presence of other topics



Example mail - Clean Up

```
Newsgroups: talk.religion.misc
From: bskendig@netcom.com (Brian Kendig)
Subject: Re: *** The list of Biblical contradictions
Message-ID: <bskendigC51CqB.K0r@netcom.com>
Organization: Starfleet Headquarters: San Francisco
References: <bskendigC50tnu.Ino@netcom.com> <7912@blue.cis.pitt.edu>
Date: Tue, 6 Apr 1993 00:15:46 GMT
Lines: 30
```

Header

```
joslin@pogo.isp.pitt.edu (David Joslin) writes:
>
>I'm curious to know what purpose people think these lists serve.
>Lists like this seem to value quantity over quality, an "argument
>from article length."  And the list you have here is of poorer
>quality than most.
```

Quotes

I agree, which is why I've asked for help with it.

The reason I'm working on this list is because I've recently had one too many Christians tell me "the Bible contains no contradictions whatsoever." They believe that it's true, and that it describes reality perfectly, and even predicts history before it happens.

Before I can carry on any sort of meaningful conversation with these people, I've got to SHOW them, with concrete evidence, that the Bible is not nearly as airtight as they thought. I hope to do that with this list.

Specifically: when I bring up the fact that Genesis contains two contradictory creation stories, I usually get blank stares or flat denials. I've never had a fundamentalist acknowledge that there are indeed two different accounts of creation.

```
--
_/_/_/  Brian Kendig                Je ne suis fait comme aucun
/_/_/  bskendig@netcom.com           de ceux que j'ai vus; j'ose croire
_/_/                        n'etre fait comme aucun de ceux qui existent.
  /  The meaning of life     Si je ne vaux pas mieux, au moins je suis autre.
 /    is that it ends.                              -- Rousseau
```

Footer

Example mail - (Possibly) Relevant Tokens for Religion Topic

I agree, which is why I've asked for help with it.

The reason I'm working on this list is because I've recently had one too many Christians tell me "the Bible contains no contradictions whatsoever." They believe that it's true, and that it describes reality perfectly, and even predicts history before it happens.

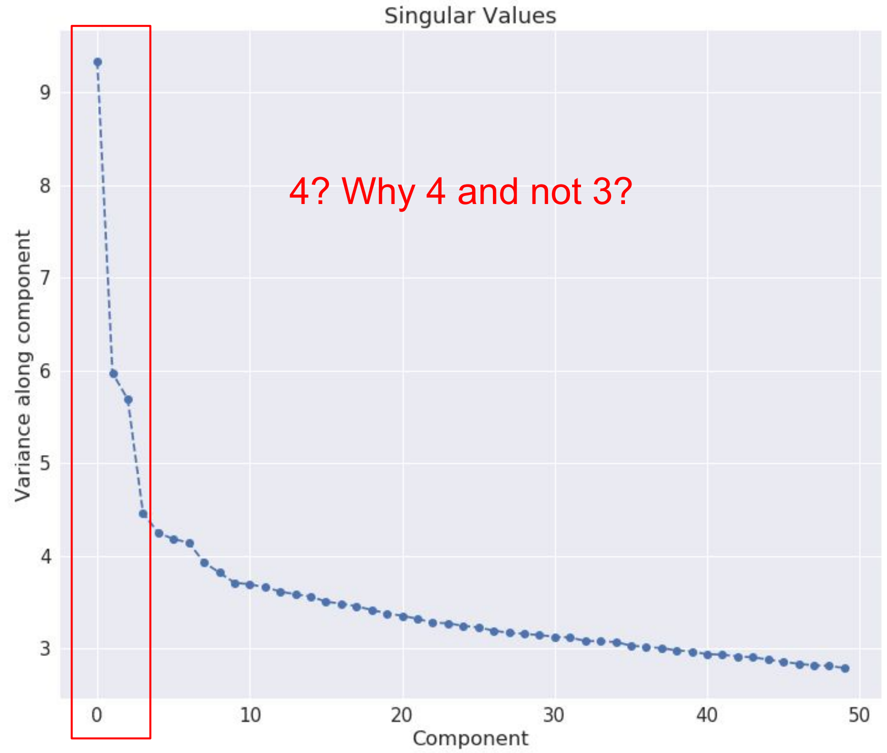
Before I can carry on any sort of meaningful conversation with these people, I've got to SHOW them, with concrete evidence, that the Bible is not nearly as airtight as they thought. I hope to do that with this list.

Specifically: when I bring up the fact that Genesis contains two contradictory creation stories, I usually get blank stares or flat denials. I've never had a fundamentalist acknowledge that there are indeed two different accounts of creation.

LSA

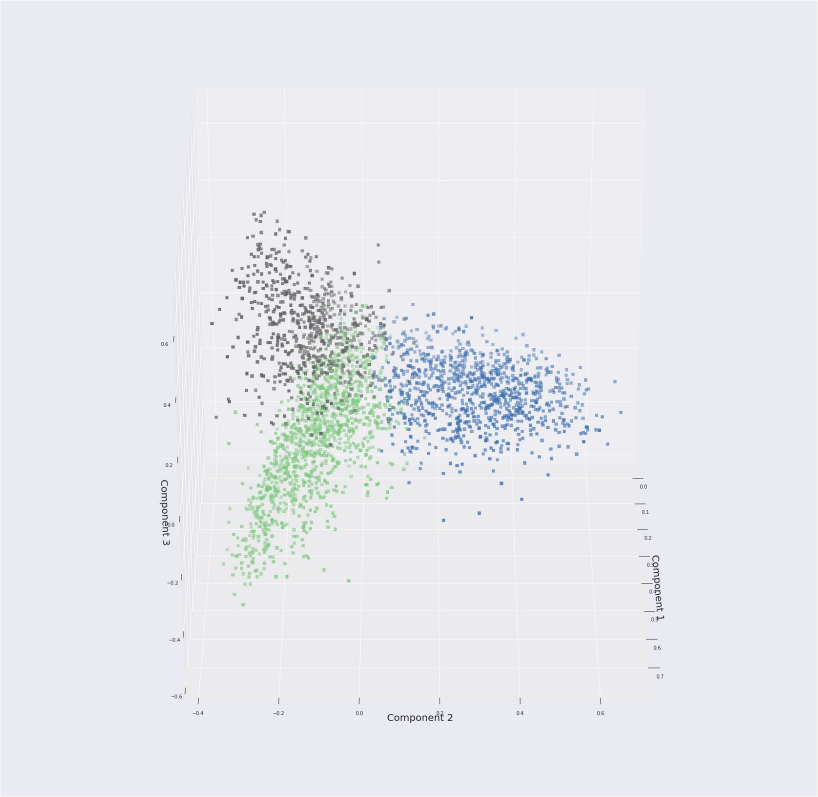
The majority of variance is explained by the first 4 components !!!

- 3000 Mail
- Each mail becomes a point in a 1000-dimensional space (TfIdf matrix)
- Reduction to 50-dimensional space(principal components) using truncated SVD
- Variance along component
- Singular values distribution imply high separation potential



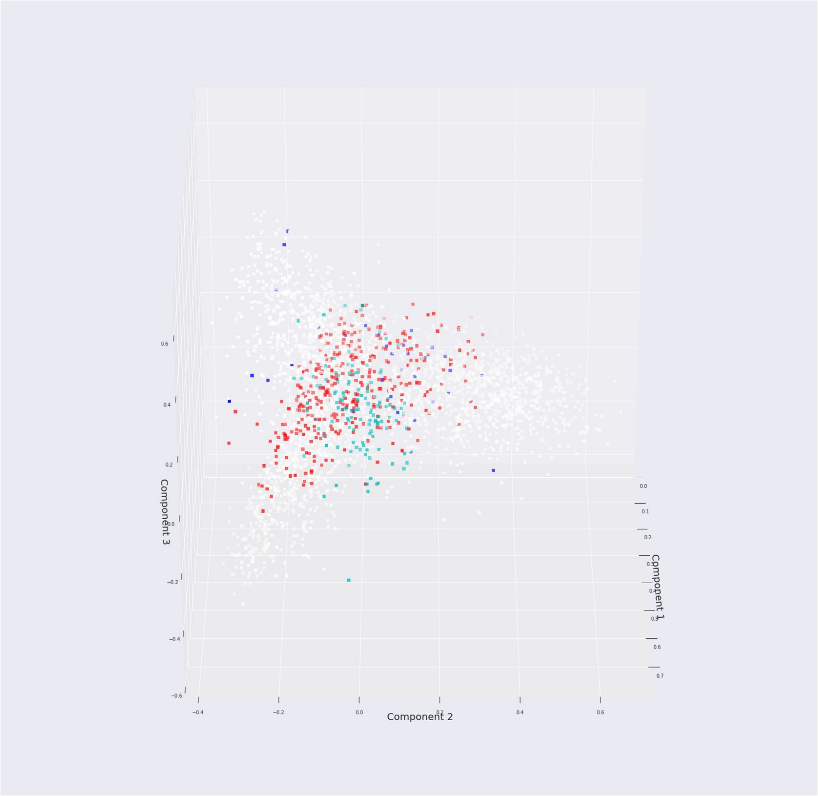
K-means Clusters

- Look for 3 clusters over 50 dimensional space (projected on the first 3 components)
- Most frequent terms per cluster:
    - Cluster 0: people god think like say know life make good jesus
    - Cluster 1: thanks graphics image know program files looking like file use
    - Cluster 2: car cars engine new like good speed ford dealer problem



K-means Clusters -results

- 513 Misclassification over 3000 e-mails:
    - Messages from rec.autos not in Cluster 2: **35.3 % (red)**
    - Messages from comp.graphics not in Cluster 1: **12.6% (cyan)**
    - Messages from talk.religion.misc not in Cluster 0: **3.4% (blue)**
    - Message in white correctly classified



Conclusions

- Dimensionality reduction through SVD improves the quality of clustering techniques for topic modeling
- Comparing the value of the components conveys informations about the topic distribution
- Extending the number of clusters beyond the number of expected categories by looking at the singular values can lead to the identification of new topics!
- **Future :**
  - Extend number of newsgroups
  - Train deep learning model on each clusters