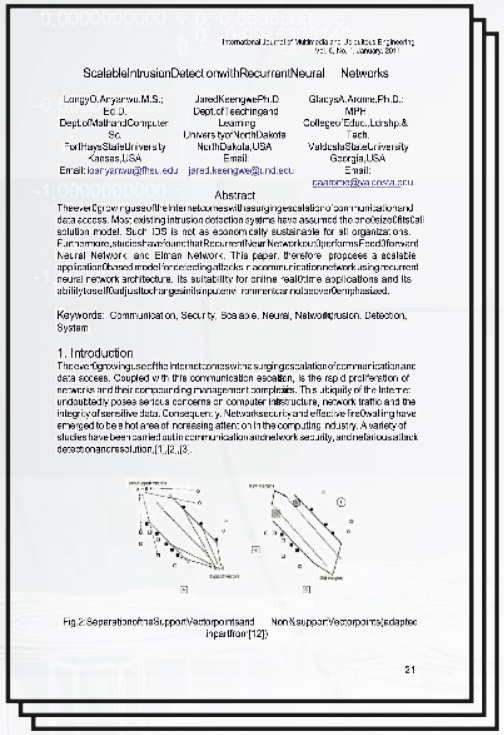
Topic modeling: a way to navigate through text collections

From texts to topics

Text documents



ScalableIntrusionDetectionwithRecurrentNeural Networks

Abstract

Keywords: Communication, Security, Scalable, Neural, Network, Intrusion, Detection, System

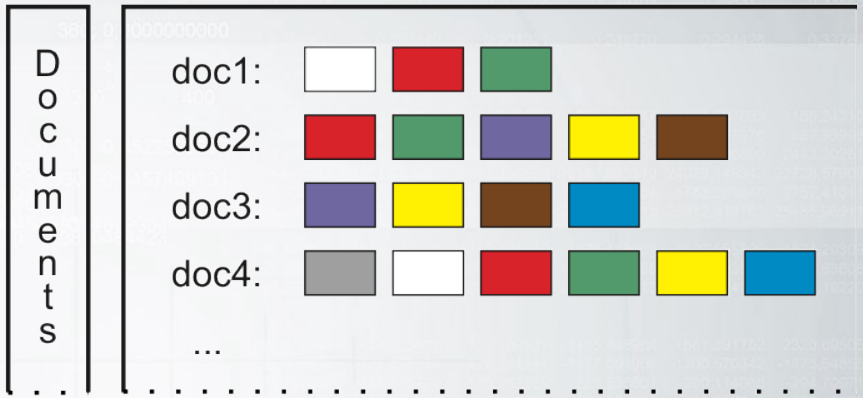
1. Introduction

Fig.2:SeperationofthaSupportVectorpointsand NonSupportVectorpoints(adaptec hpartition'[12])

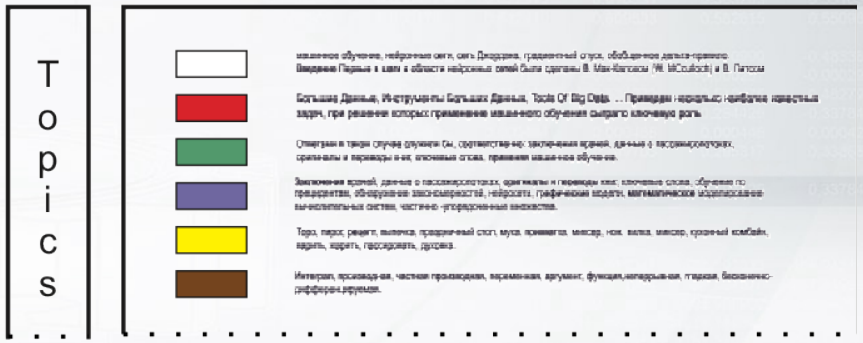
21

Topic Modeling

Topics of documents



Words and keyphrases of topics



The formal task

Given:

- Collection of texts as bags-of-words:

  n_{wd} is a count of the word w in the document d

Find:

- Probabilities of word in topics:

  $\phi_{wt} = p(w|t)$

- Probabilities of topics in documents:

  $\theta_{td} = p(t|d)$

The formal task

Given:

- Collection of texts as bags-of-words:

  n_{wd} is a count of the word w in the document d

Find:

- Probabilities of word in topics:

  $\phi_{wt} = p(w|t)$ ⟵ **Definition of a topic!**

- Probabilities of topics in documents:

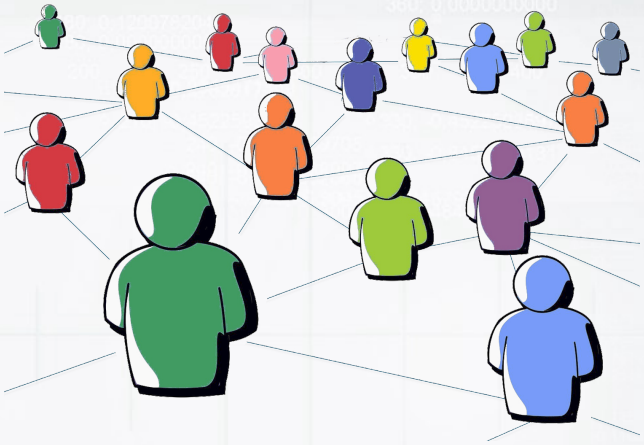
  $\theta_{td} = p(t|d)$

Where do we need that?

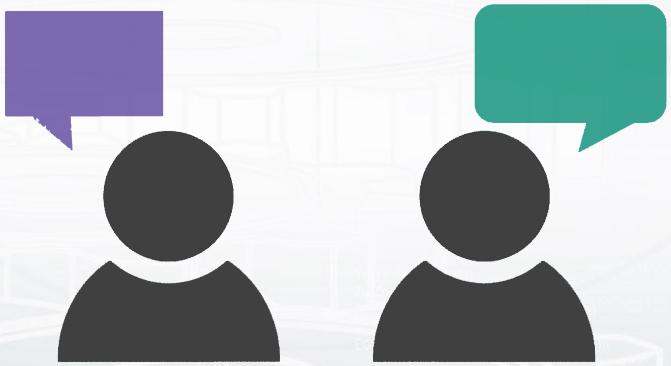
Exploration and navigation through large text collections

Where do we need that?

- Social network analysis



- Dialogue manager in chat-bots



Why do we need it?

Topic models provide hidden semantic representation of texts.

Many more applications:

- Categorization and classification of texts
- Document segmentation and summarization
- News flows aggregation and analysis
- Recommender systems
- Image captioning
- Bioinformatics (genome annotation)
- Exploratory search
- …

Generative model of texts

Probabilistic Latent Semantic Analysis (PLSA):

$$p(w|d) = \sum_{t \in T} p(w|t,d)\, p(t|d) = \sum_{t \in T} p(w|t)\, p(t|d)$$

Notation:

- w – *word*
- d – *document*
- t – *topic*

Generative model of texts

Probabilistic Latent Semantic Analysis (PLSA):

$$p(w|d) = \sum_{t \in T} p(w|t,d)\, p(t|d) = \sum_{t \in T} p(w|t)\, p(t|d)$$

Law of total probability

$$p(w) = \sum_{t \in T} p(w|t)\, p(t)$$

Notation:

- w – *word*
- d – *document*
- t – *topic*

Generative model of texts

Probabilistic Latent Semantic Analysis (PLSA):

$$p(w|d) = \sum_{t \in T} p(w|t,d)\, p(t|d) = \sum_{t \in T} p(w|t)\, p(t|d)$$

Law of total probability

$$p(w) = \sum_{t \in T} p(w|t)\, p(t)$$

Assumption of conditional independence

$$p(w|t,d) = p(w|t)$$

Notation:

- w – *word*
- d – *document*
- t – *topic*

Generative model of texts

Topics

gene	0.04
dna	0.02
genetic	0.01
. . .	

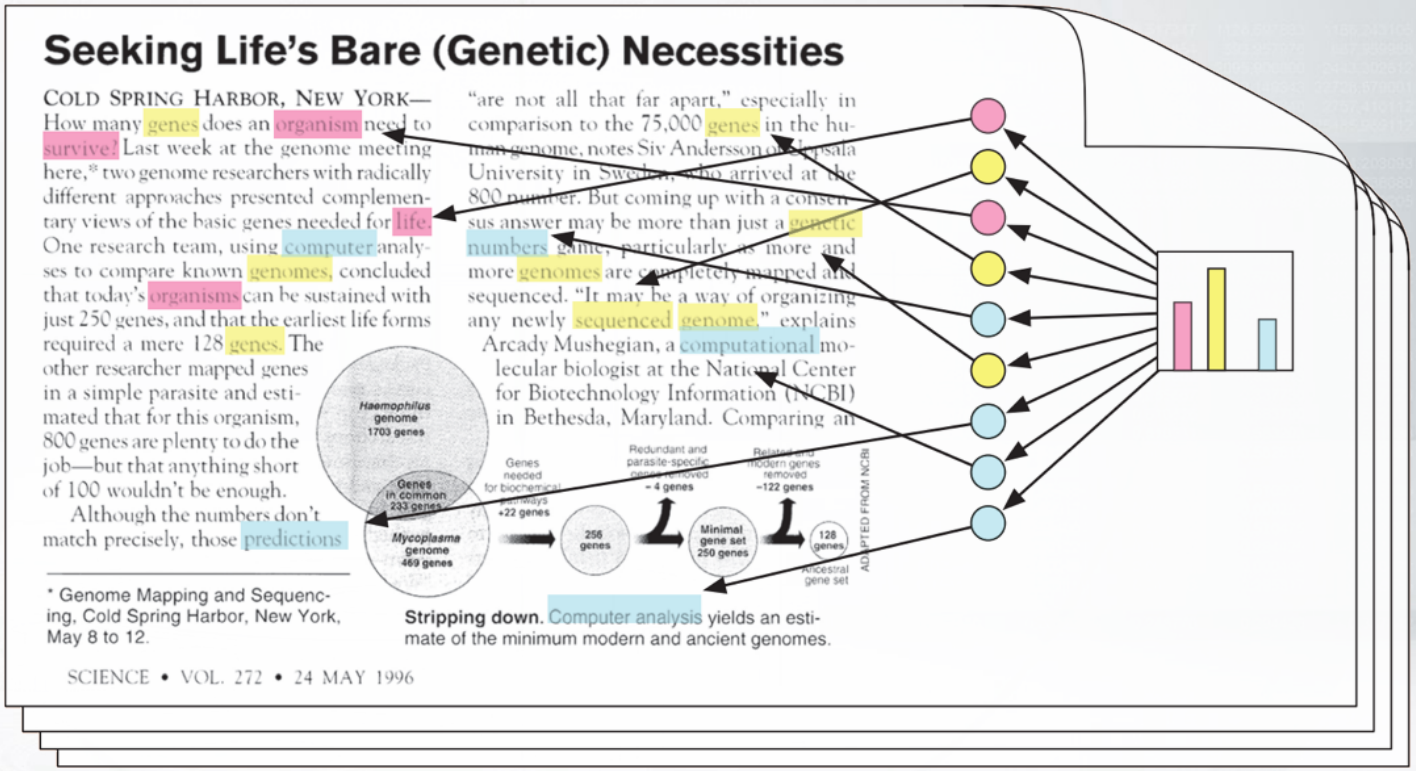
life	0.02
evolve	0.01
organism	0.01
. . .	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02
computer	0.01
. . .	

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

David Blei, Probabilistic topic models, 2012.

Matrix way of thinking

Probabilistic Latent Semantic Analysis:

$$p(w|d) = \sum_{t \in T} p(w|t)\, p(t|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$$