Text Analytics

Daniel Archambault

Documents: What to Do?

- Sometimes your data is a collection of documents
  - social media posts
  - collections of documents (panama papers)
- Text doesn't behave quite the same
- But still need to analyse at scale

An Analysis Approach: Similarity

Given a document collection, present a visualisation that represents the degree of similarity between documents.

- Discard grammar and use only vocabulary for scalability
- Documents that are similar should be close to each other
- Documents that are not similar should be further away
- Sounds familiar...

One way to model: Bag of Words

my document: *Aardvarks play with zebra...*

$$\text{mydocument} = \begin{array}{c} \textit{Aardvarks} \\ \ldots \\ \textit{play} \\ \ldots \\ \textit{Tokyo} \\ \ldots \\ \textit{zebra} \end{array} \begin{bmatrix} 1 \\ \ldots \\ 1 \\ \ldots \\ 0 \\ \ldots \\ 1 \end{bmatrix}$$

- Model the document as a collection of words and count frequencies
- Transform the data into a high dimensional data set
- In this data set, documents will be close if vocabulary is the same
- Get rid of words with little meaning (a, an, the...) – stop words
- Need an appropriate distance measure...

TF-IDF: Not all words are equal

- Measurement of how unique a word is given a collection of documents

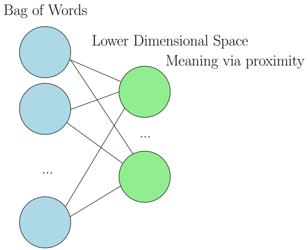
- term-frequency $tf = \frac{w}{n}$: n words in document, w number of times word occurs
- inverse document frequency $idf = \log \frac{N}{N_w}$: N number of documents, N_w number of documents with the word counted in w
- TF-IDF : $tf \cdot idf$ measure of how unique word is in documents
- Can be used for normalisation in bag of words models

word2vec

- Bag of Words can produce very high dimensional spaces
- Many of these dimensions do not contribute much
- Can we collapse the dimensionality down so that we have similar information but fewer dimensions?
- Yes we can through machine learning

word2vec

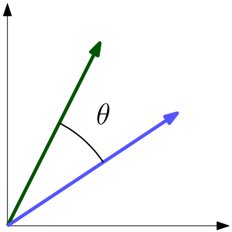
- Compute an embedding of words in a lower dimensional space
- Words used in the same way are in similar areas of the lower dimensional space



- train neural network to convert vectors to lower dimensional space
- context can be document or collection of sentences

Cosine Similarity

- Turns out that straight line distance is not good in these spaces
- But, angle between vectors is a good measure



- Take the cosine of the angle to make between $[0, 1]$
  - $\cos\theta = 1$ - vocabulary (vectors) is the same
  - $\cos\theta = 0$ - vocabulary (vectors) are orthogonal
- Problem, how to measure cosine in d dimensions?

Dot product gives you $\cos\theta$

- The dot product gives you $\cos\theta$
- You can express in d dimensions

$$\vec{a} \cdot \vec{b} = ||\vec{a}||||\vec{b}|| \cos\theta$$

$$\cos\theta = \frac{\sum_{i=1}^{d} a_i b_i}{||\vec{a}||||\vec{b}||}$$

- measure close to zero for dissimilar documents
- measure close to one for similar documents

Similar documents, Different documents

a: *Aardvarks play with zebra...*
b: *Tokyo and Olympics...*
c: *No aardvarks or zebra in Tokyo ...*

$$\vec{a}\cdot\vec{b} = \begin{array}{c} \textit{Aardvarks} \\ \ldots \\ \textit{Olympics} \\ \ldots \\ \textit{play} \\ \ldots \\ \textit{Tokyo} \\ \ldots \\ \textit{zebra} \end{array} \begin{bmatrix} 1 \\ \ldots \\ 0 \\ \ldots \\ 1 \\ \ldots \\ 0 \\ \ldots \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ \ldots \\ 1 \\ \ldots \\ 0 \\ \ldots \\ 1 \\ \ldots \\ 0 \end{bmatrix} = 0 \qquad \vec{a}\cdot\vec{c} = \begin{array}{c} \textit{Aardvarks} \\ \ldots \\ \textit{Olympics} \\ \ldots \\ \textit{play} \\ \ldots \\ \textit{Tokyo} \\ \ldots \\ \textit{zebra} \end{array} \begin{bmatrix} 1 \\ \ldots \\ 0 \\ \ldots \\ 0 \\ \ldots \\ 0 \\ \ldots \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \ldots \\ 0 \\ \ldots \\ 0 \\ \ldots \\ 1 \\ \ldots \\ 1 \end{bmatrix} \approx 1$$

How to convert distances to two dimensions?

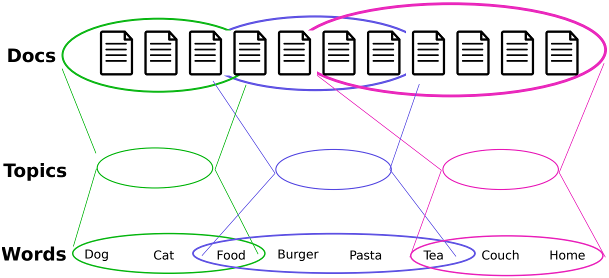
- The cosine similarity can be computed for every pair of vectors
- Loaded into a distance matrix
- Use any dimensionality reduction technique
  - e.g. multidimensional scaling will work

Latent Dirichlet allocation (LDA)

- How to group documents together?
- Each document can belong to many topics
- Topic description should be minimal set of words
- Related documents will have similar words
- How to link documents through words efficiently?

Latent Dirichlet allocation (LDA)

- Can be seen as a fuzzy clustering
- Topics are defined as a distribution of words
- Words are mapped to topics
- Documents are mapped to topics
- Mapping is done probabilistically



Overview

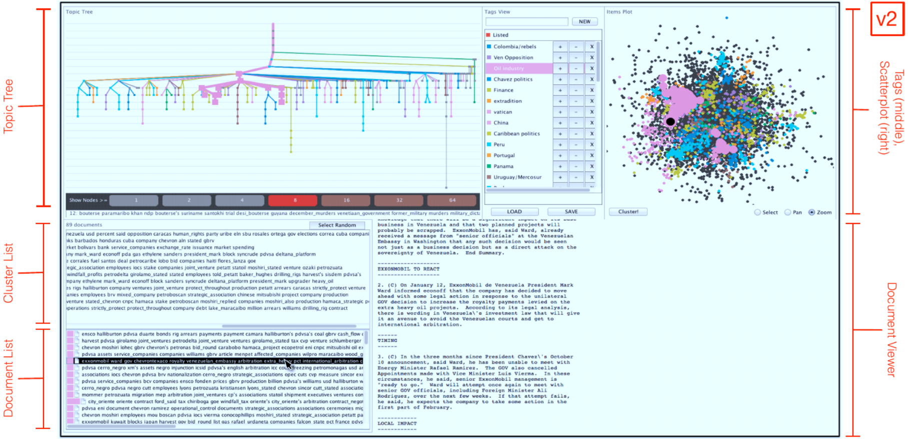
The standard approach:

- Encode the documents somehow (TF-IDF)
- Use cosine distance for similarity
- Cluster the documents into topics (hierarchical)
- Visualization to browse topics/documents

M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists.

IEEE Transactions on Visualization and Computer Graphics, 20(12):2271–2280, 2014

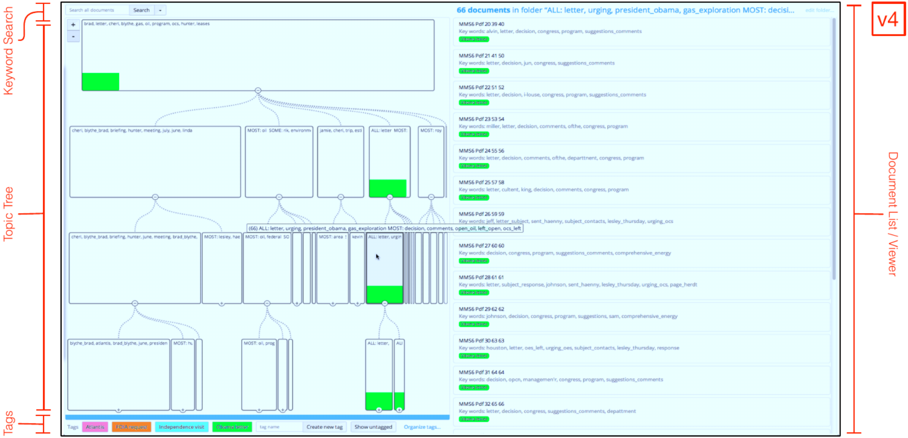
Overview



M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists.

IEEE Transactions on Visualization and Computer Graphics, 20(12):2271–2280, 2014

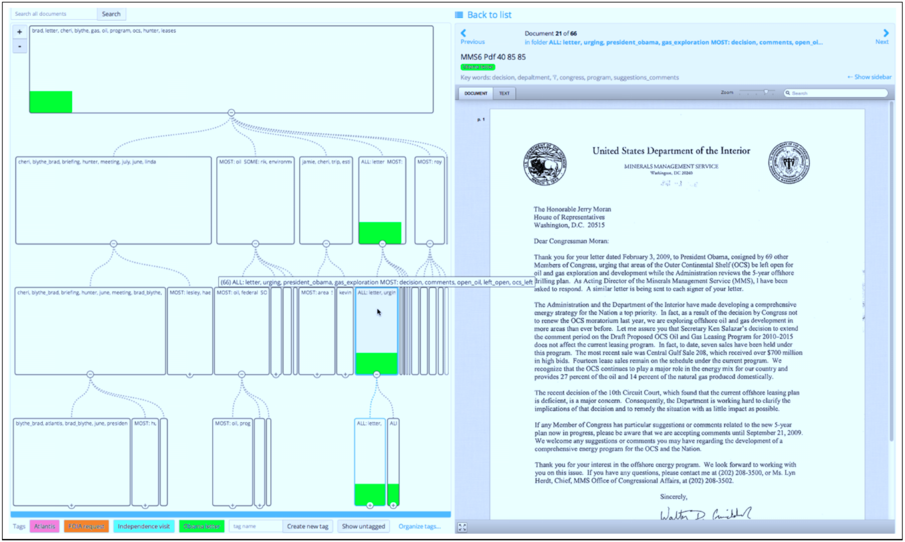
Overview



M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists.

IEEE Transactions on Visualization and Computer Graphics, 20(12):2271–2280, 2014.

Overview


Search all documents
Search
brad, letter, cheri, blythe, gas, oil, program, ocs, hunter, leases
cheri, blythe_brad, briefing, hunter, meeting, july, june, linda
MOST: oil SOME: rik, environm
jamie, cheri, trip, esti
ALL: letter MOST:
MOST: roy
(66) ALL: letter, urging, president_obama, gas_exploration MOST: decision, comments, open_oil, left_open, ocs_left
cheri, blythe_brad, briefing, hunter, june, meeting, brad_blythe,
MOST: lesley, hae
MOST: oil, federal SO
MOST: area
kevin
ALL: letter, urgin
blythe_brad, atlantis, brad_blythe, june, presiden
MOST: hu
MOST: oil, prog
ALL: letter,
AU
Tags
Atlantis
FOIA request
Independence visit
Obama letter
tag name
Create new tag
Show untagged
Organize tags...
Back to list
Previous
Document 21 of 66
in folder ALL: letter, urging, president_obama, gas_exploration MOST: decision, comments, open_oil...
Next
MMS6 Pdf 40 85 85
Obama letter
Key words: decision, depaltment, 'i', congress, program, suggestions_comments
← Show sidebar
DOCUMENT
TEXT
Zoom
Search
p. 1
United States Department of the Interior
MINERALS MANAGEMENT SERVICE
Washington, DC 20240
The Honorable Jerry Moran
House of Representatives
Washington, D.C. 20515
Dear Congressman Moran:
Thank you for your letter dated February 3, 2009, to President Obama, cosigned by 69 other Members of Congress, urging that areas of the Outer Continental Shelf (OCS) be left open for oil and gas exploration and development while the Administration reviews the 5-year offshore drilling plan. As Acting Director of the Minerals Management Service (MMS), I have been asked to respond. A similar letter is being sent to each signer of your letter.
The Administration and the Department of the Interior have made developing a comprehensive energy strategy for the Nation a top priority. In fact, as a result of the decision by Congress not to renew the OCS moratorium last year, we are exploring offshore oil and gas development in more areas than ever before. Let me assure you that Secretary Ken Salazar's decision to extend the comment period on the Draft Proposed OCS Oil and Gas Leasing Program for 2010-2015 does not affect the current leasing program. In fact, to date, seven sales have been held under this program. The most recent sale was Central Gulf Sale 208, which received over $700 million in high bids. Fourteen lease sales remain on the schedule under the current program. We recognize that the OCS continues to play a major role in the energy mix for our country and provides 27 percent of the oil and 14 percent of the natural gas produced domestically.
The recent decision of the 10th Circuit Court, which found that the current offshore leasing plan is deficient, is a major concern. Consequently, the Department is working hard to clarify the implications of that decision and to remedy the situation with as little impact as possible.
If any Member of Congress has particular suggestions or comments related to the new 5-year plan now in progress, please be aware that we are accepting comments until September 21, 2009. We welcome any suggestions or comments you may have regarding the development of a comprehensive energy program for the OCS and the Nation.
Thank you for your interest in the offshore energy program. We look forward to working with you on this issue. If you have any questions, please contact me at (202) 208-3500, or Ms. Lyn Herdt, Chief, MMS Office of Congressional Affairs, at (202) 208-3502.
Sincerely,

Serendip

- What if documents are in more than one topic?
- How do topics correspond to elements in the text?
- Which words are linked to which documents?

Solution

- LDA for topic modelling
- Different views to show different aspects of topics, documents, and words

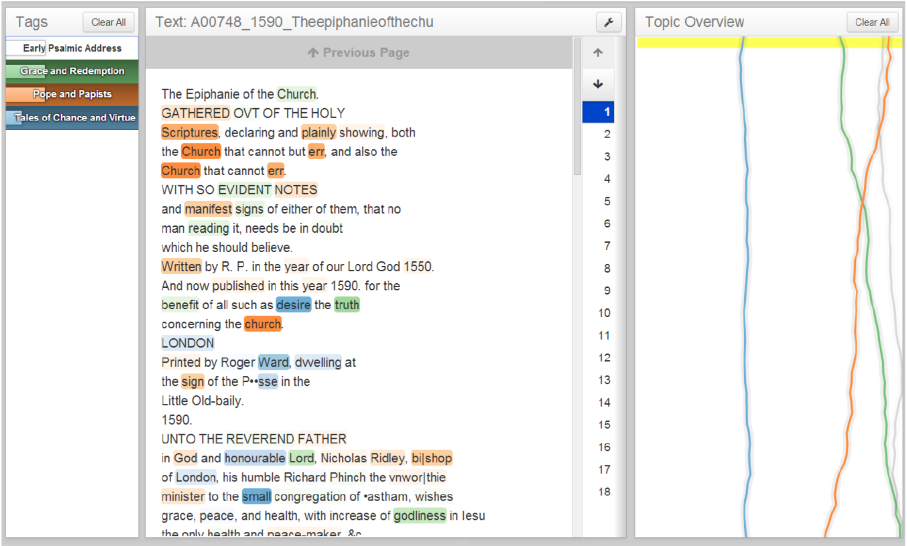
E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text corpora.

In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 173–182, October 2014

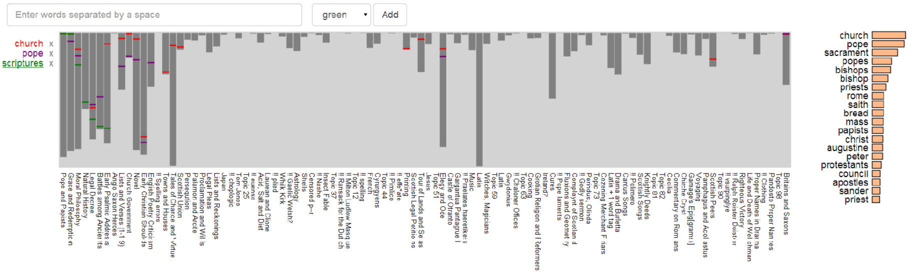
Serendip

E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text

Serendip


Tags
Clear All
Early Psalmic Address
Grace and Redemption
Pope and Papists
Tales of Chance and Virtue
Text: A00748_1590_Theepiphanieofthechu
Previous Page
The Epiphanie of the Church.
GATHERED OVT OF THE HOLY
Scriptures, declaring and plainly showing, both
the Church that cannot but err, and also the
Church that cannot err.
WITH SO EVIDENT NOTES
and manifest signs of either of them, that no
man reading it, needs be in doubt
which he should believe.
Written by R. P. in the year of our Lord God 1550.
And now published in this year 1590. for the
benefit of all such as desire the truth
concerning the church.
LONDON
Printed by Roger Ward, dwelling at
the sign of the P••sse in the
Little Old-baily.
1590.
UNTO THE REVEREND FATHER
in God and honourable Lord, Nicholas Ridley, bi|shop
of London, his humble Richard Phinch the vnwor|thie
minister to the small congregation of •astham, wishes
grace, peace, and health, with increase of godliness in Iesu
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Topic Overview
Clear All

Serendip



E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text corpora.

In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 173–182, October 2014

Further reading

If you're interested…

- Franz Wanner, Andreas Stoffel, Dominik Jäckle, Bum Chul Kwon, Andreas Weiler, and Daniel A. Keim. State-of-the-art report of visual analysis for event detection in text data streams. In R. Borgo, R. Maciejewski, and I. Viola, editors, *EuroVis - STARs*, pages 125–139, Swansea, UK, 2014
- Kostiantyn Kucher, Carita Paradis, and Andreas Kerren. The state of the art in sentiment visualization. *Computer Graphics Forum*, 37(1):71–96, 2018

[AKV+14] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text corpora. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 173–182, October 2014.

[BISM14] M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2271–2280, 2014.

[KPK18] Kostiantyn Kucher, Carita Paradis, and Andreas Kerren. The state of the art in sentiment visualization. *Computer Graphics Forum*, 37(1):71–96, 2018.

[WSJ+14] Franz Wanner, Andreas Stoffel, Dominik Jäckle, Bum Chul Kwon, Andreas Weiler, and Daniel A. Keim. State-of-the-art report of visual analysis for event detection in text data streams. In R. Borgo, R. Maciejewski, and I. Viola, editors, *EuroVis - STARs*, pages 125–139, Swansea, UK, 2014.