Text Similarity

Class Algorithmic Methods of Data Mining
Program M. Sc. Data Science
University Sapienza University of Rome
Semester Fall 2015
Lecturer Carlos Castillo http://chato.cl/

Sources:

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. Sections 6.2, 6.3.

Google catholic

Web News Maps Images Books More Search tools

Catholic News A...

Catholic California governor: I considered the perspective…
Washington Post - 12 hours ago
California Gov. Jerry Brown (D), a lifelong **Catholic** and former Jesuit seminarian, signed a law Monday legalizing physician-assisted suicide in ...

Despite objections, Calif. governor signs assisted-suicide bill
Catholic News Agency - 11 hours ago

Explore in depth (1,015 more articles)

Is the **Catholic** Church Ready to Make Changes on Homo…
The Atlantic - 14 hours ago
Within the last three days, two gay **Catholic** priests have been fired from their positions because of their relationships with adult men. Over the ...

Conservatives Launch Antigay Salvo as **Catholic** Bishops' Meeting ...
Advocate.com - 12 hours ago

Pope Francis opens Roman **Catholic** synod amid gay row
BBC News - Oct 4, 2015

Does Pope Francis fear God? On the Synod of the Family and the ...
Opinion - The Week Magazine - Oct 5, 2015

Will the Vatican make it easier to be a divorced **Catholic**?
In-Depth - Christian Science Monitor - Oct 4, 2015

Pope Francis dismisses "passing fads" on marriage
Opinion - CBS News - Oct 4, 2015

Advocate.com BBC News ABC Online Irish Indepen... Christian Sci... National Cat...

Explore in depth (1,495 more articles)

Why are these similar?

Why are these different?

Is the Catholic Church Ready to Make Changes on Homo…
The Atlantic - 14 hours ago
Within the last three days, two gay **Catholic** priests have been fired from their positions because of their relationships with adult men. Over the ...

Conservatives Launch Antigay Salvo as Catholic Bishops' Meeting ...
Advocate.com - 12 hours ago

Pope Francis opens Roman Catholic synod amid gay row
BBC News - Oct 4, 2015

Does Pope Francis fear God? On the Synod of the Family and the ...
Opinion - The Week Magazine - Oct 5, 2015

Will the Vatican make it easier to be a divorced Catholic?
In-Depth - Christian Science Monitor - Oct 4, 2015

Pope Francis dismisses "passing fads" on marriage
Opinion - CBS News - Oct 4, 2015


Advocate.com BBC News ABC Online Irish Indepen… Christian Sci… National Cat…

Explore in depth (1,495 more articles)


USA TODAY

Catholic Church struggles to retain LGBT-supporting mille…
The Daily Cardinal - 8 hours ago
During the video conference, Charamsa called upon Pope Francis to revise the **Catholic** doctrine on homosexuality. The current doctrine ...

Vatican fires gay priest on eve of Catholic bishops meeting
USA TODAY - Oct 3, 2015

Catholic Priest Comes Out, Reveals Gay Boyfriend, Gets Fired by ...
Mediaite - Oct 4, 2015

Explore in depth (621 more articles)

Why are these similar?

Why are these different?

Various levels of text similarity

- String distance (e.g. edit distance)
- Lexical overlap
- **Vector space model**
  - **A simple model of text similarity, introduced by Salton et al. in 1975**
- Usage of semantic resources
- Automatic reasoning/understanding
- AI-complete text similarity

Running example

- Q: "gold silver truck"
- D1: "Shipment of gold damaged in a fire"
- D2: "Delivery of silver arrived in a silver truck"
- D3: "Shipment of gold arrived in a truck"

Which document is more similar to Q?

Bag of Words Model: Binary vectors

- First, normalize (in this case, lowercase)
- Second, compute vocabulary and sort
  - a arrived damaged delivery fire gold in of shipment silver truck

	a	arrived	damaged	delivery	fire	gold	in	of	shipment	silver	truck
Shipment of gold damaged in a fire	1	0	1	0	1	1	1	1	1	0	0
Shipment of gold arrived in a truck	1	1	0	0	0	1	1	1	1	0	1

Distance

	a	arrived	damaged	delivery	fire	gold	in	of	shipment	silver	truck
D1.Shipment of gold damaged in a fire	1	0	1	0	1	1	1	1	1	0	0
D2. Shipment of gold arrived in a truck	1	1	0	0	0	1	1	1	1	0	1

- Similarity between D1 and D2
  - <v1,v2> = 5

What are the shortcomings of this method?

How important is a term?

- Common terms such as “a”,”in”,”of” don't say much
- Rare terms such as “gold”,”truck” say more
- Document Frequency of term *t*
  - Number of documents containing a term
  - *DF(t)*
- Inverse document frequency of term *t*
  - $IDF(t) = \log\left(\frac{|D|}{DF(t)}\right)$

Example

- D1: "Shipment of gold damaged in a fire"
- D2: "Delivery of silver arrived in a silver truck"
- D3: "Shipment of gold arrived in a truck"

- |D| = 3
- IDF(“gold”) = ?
- IDF(“a”) = ?
- IDF(“silver”) = ?

Example

- D1: "Shipment of gold damaged in a fire"
- D2: "Delivery of silver arrived in a silver truck"
- D3: "Shipment of gold arrived in a truck"

- |D| = 3
- IDF(“gold”) = log(3 / 2) = 0.176 (using $\log_{10}$)
- IDF(“a”) = log(3 / 3) = 0.000
- IDF(“silver”) = log(3 / 1) = 0.477

Term frequency

- Term frequency(doc,term) = TF(doc,term)
  - Number of times the term appears in a document
- If a document contains many occurrences of a word, that word is important for the document

TF("Delivery of **silver** arrived in a **silver** truck", "silver") = 2

TF("Delivery of silver **arrived** in a silver truck","arrived") = 1

Document vectors

- Di,j corresponds to document i, term j
- Di,j = TF(i,j) x IDF(j)

Exercise:
Write the document vectors for all 3 documents

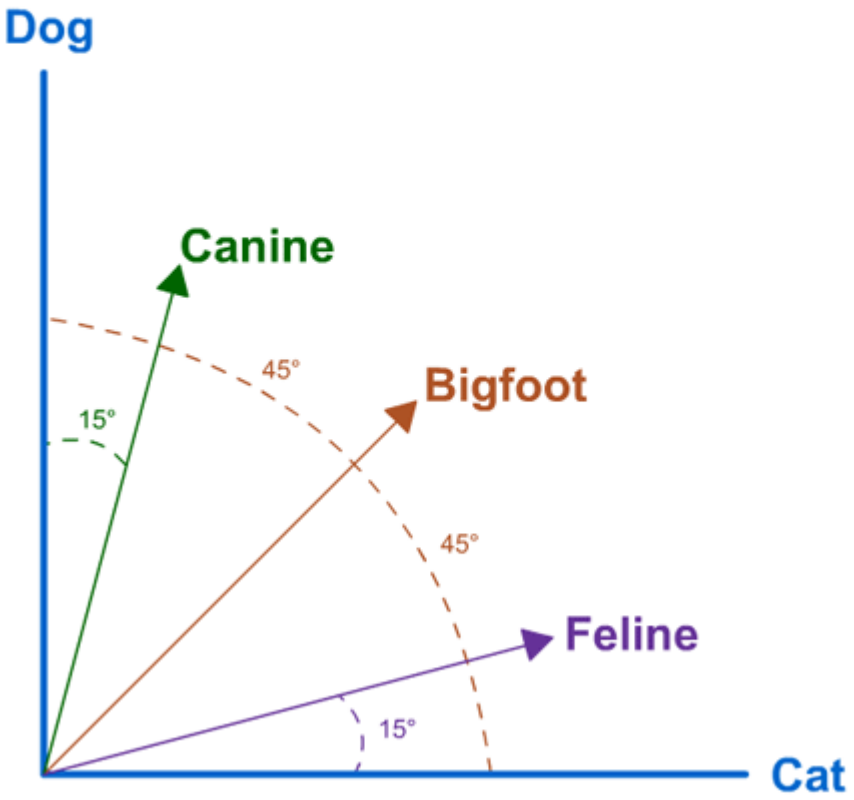
- *D1: "Shipment of gold damaged in a fire"*
- *D2: "Delivery of silver arrived in a silver truck"*
- *D3: "Shipment of gold arrived in a truck"*

Verify: D1,3 = 0.477; D2,10=0.954

Answer:

http://chato.cl/2015/data_analysis/exercise-answers/textdistance_exercise_01_answer.pdf

Computing similarity



- Each document is a vector in the positive quadrant
- The cosine of the angle between vectors is their similarity

$$\cos(D_1, D_2) = \frac{D_1 \cdot D_2}{|D_1||D_2|}$$

Image: https://moz.com/blog/lda-and-googles-rankings-well-correlated

What is the best document?

- Q = “gold silver truck”

Exercise:

- *Write the TF·IDF vector for Q*
- *Compute D1·Q, D2·Q, D3·Q (do not normalize)*
- *Verify you got these numbers (in a different ordering): { 0.062, 0.031, 0.486 }*
- *What is the best document?*

Answer:
http://chato.cl/2015/data_analysis/exercise-answers/textdistance_exercise_01_answer.pdf

Pros/Cons

- We are losing information
  - Sentence structure, proximity of words, ordering of the words
    - *How could we keep this?*
  - Capitalization and everything we lost during normalization
    - *How could we keep this?*
- But
  - It's really fast
  - It works in practice
  - It can be extended, e.g. different weighting schemes

R.it PALERMO

Home | Cronaca | Sport | Foto | Ristoranti | Annunci Loca

Il buco nel bilancio della Regione: audizioni alla Corte dei Conti

Davanti ai giudici contabili gli assessori Baccei e Gucciardi insieme al ragioniere generale e altri dirigenti. Gucciardi: "Le assunzioni si faranno, non siamo più Regione canaglia"

Lo leggo dopo | 06 ottobre 2015

Notifiche

Sono iniziate davanti alla Corte dei conti, le audizioni di dirigenti e assessori della Regione siciliana convocati dal Presidente Maurizio Graffeo, per procedere alla verifica dello stato dei conti pubblici della Regione prima della chiusura dell'esercizio finanziario 2015. La Corte dei conti ha deciso di convocare a sezioni riunite l'assessore all'economia Alessandro Baccei, l'assessore alla Salute Baldo Gucciardi, il ragioniere generale della Regione Salvatore Sammartano, e i dirigenti generali della Programmazione, Vincenzo Falgares, dell'Agricoltura, Giovanni Bologna e della Formazione, Gianni Silvia, cioè i responsabili dell'impiego dei fondi europei. Dai rilievi fatti dalla Corte dei Conti potrebbe derivare anche un aggiustamento dei documenti finanziari. Prima delle audizioni gli assessori Baccei e Gucciardi, interpellati dai giornalisti, non hanno voluto commentare le audizioni.

Al centro della audizione lo stato dei conti regionali: si parla di un ammanco da 400 milioni di euro. Ma, soprattutto, i giudici chiederanno, "se risponde al vero quanto letto su giornali riguardo alle 5 mila assunzioni nella Sanita'" che sarebbero non sostenibili economicamente dalla Sicilia: "Siamo ancora in

Weighting schemes

Most documents have some structure

This structure allows us to do something better than TF

What would you do?