NLP

Text similarity

Introduction

Text Similarity

- People can express the same concept (or related concepts) in many different ways. For example, “the plane leaves at 12pm” vs “the flight departs at noon”
- Text similarity is a key component of Natural Language Processing
- If the user is looking for information about cats, we may want the NLP system to return documents that mention kittens even if the word “cat” is not in them.
- If the user is looking for information about “fruit dessert”, we want the NLP system to return documents about “peach tart” or “apple cobbler”.
- A speech recognition system should be able to tell the difference between similar sounding words like the “Dulles” and “Dallas” airports.
- This set of lectures will teach you how text similarity can be modeled computationally.

Human Judgments of Similarity

tiger	cat	7.35
tiger	tiger	10.00
book	paper	7.46
computer	keyboard	7.62
computer	internet	7.58
plane	car	5.77
train	car	6.31
telephone	communication	7.50
television	radio	6.77
media	radio	7.42
drug	abuse	6.85
bread	butter	6.19
cucumber	potato	5.92

[Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin, "Placing Search in Context: The Concept Revisited", ACM Transactions on Information Systems, 20(1):116-131, January 2002]

Human Judgments of Similarity

delightful	wonderful	A	8.65
modest	flexible	A	0.98
clarify	explain	V	8.33
remind	forget	V	0.87
get	remain	V	1.6
realize	discover	V	7.47
argue	persuade	V	6.23
pursue	persuade	V	3.17
plane	airport	N	3.65
uncle	aunt	N	5.5
horse	mare	N	8.33

[SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. 2014. Felix Hill, Roi Reichart and Anna Korhonen. Preprint pubslished on arXiv. arXiv:1408.3456]

Automatic Similarity Computation

spain	0.679
belgium	0.666
netherlands	0.652
italy	0.633
switzerland	0.622
luxembourg	0.610
portugal	0.577
russia	0.572
germany	0.563
catalonia	0.534

- Words most similar to “France”
- Computed using “word2vec”
- [Mikolov et al. 2013]

Types Of Text Similarity

- Many types of text similarity exist:
  - Morphological similarity (e.g., respect–respectful)
  - Spelling similarity (e.g., theater–theatre)
  - Synonymy (e.g., talkative–chatty)
  - Homophony (e.g., raise–raze–rays)
  - Semantic similarity (e.g., cat–tabby)
  - Sentence similarity (e.g., paraphrases)
  - Document similarity (e.g., two news stories on the same event)
  - Cross-lingual similarity (e.g., Japan–Nihon)

NLP