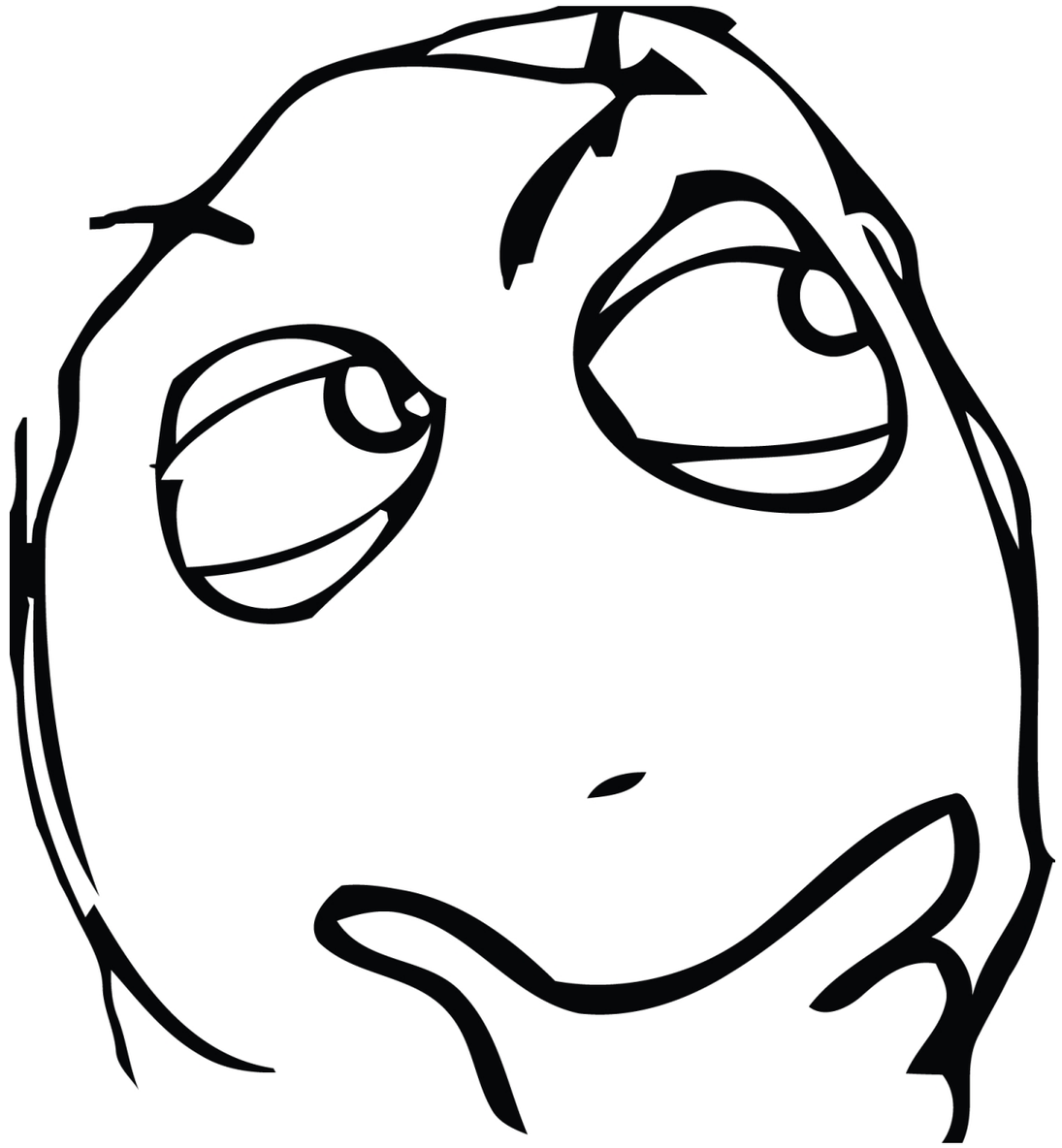
Text Summarization

Natural Language Processing Laboratory
2017/05/02

What is Summarization?



Google Search

Google | Gates microsoft 2000

Search | About 14,200,000 results (0.32 seconds)

Web
Images
Maps
Videos
News
More

Nashik, Maharashtra
Change location

Enter "Generation I" by Bill **Gates** - **Microsoft**
www.**microsoft**.com/presspass/ofnote/03-00instructor.mspx
This article appeared in the March **2000** issue of magazine. by Bill **Gates**. We are living in a truly fascinating time of change and promise. Within just a few years, **...**

Bill **Gates** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Bill_**Gates**
Gates stepped down as chief executive officer of **Microsoft** in January **2000**. He remained as chairman and created the position of chief software architect.

United States v. **Microsoft** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/United_States_v._**Microsoft**
Microsoft Chairman Bill **Gates** was called "evasive and nonresponsive" by a **...** On April 3, **2000**, he issued his conclusions of law, according to which **Microsoft ...**

More example

- 一天搞懂深度學習
- 三分鐘陳興國文懶人包
- 5分鐘看完XXX

Why We Need?

- Explosive quantity of information
- Everyone has only 24 hours every day
- Massive Information overload
- To save time and money

What We Need?

- The most important points of the document
- People, Event, Time, Place, Statistics, Consequence, …

To take an original article, understand it and pack it neatly into a nutshell without loss of substance or clarity presents a challenge which many have felt worth taking up for the joys of achievement alone. These are the characteristics of an art form

–Ashworth

Various Forms of Summarization

- Picture Storytelling
- Facebook Year Review
- NBA highlights
- …

Text Summarization

- Reduce a text document
- While retaining the most important points of the original document
- A shorter and concise version of the original document

Form of Summarization

- Extractive
  - Select sentences, passages from the original text
- Abstractive
  - Concisely paraphrase the information content
  - Require natural language understanding and generation
  - Become more active these days

Book Review?

An innocent hobbit of The Shire journeys with eight companions to the fires of Mount Doom to destroy the One Ring and the dark lord Sauron forever.

Movie Trailer?

How about 谷阿莫?

Dimension

- Single
  - Key ideas of single document
- Multi-document
  - Extraction of information from multiple texts about the same topic
  - Summary report

Query-Specific vs. Generic

- A generic summary makes no assumption about the reader's interests
- Query-specific summaries are specialized for a single information need, the query
- Summarization is much easier if we have a description of what the user wants

Top-down vs Bottom-up Summarization

- Top-down

- "I know what I want; give me what I ask for".

- User needs: only certain types of information

- Particular criteria of interest for focused search

- Templates, term lists

- Bottom-up

- "I'm curious to know what's there in the text".

- User needs: anything that's important

- Generic information metrics

- Connectedness of sentences, words frequencies

Genre

- Some genres are easy to summarize
  - Newswire stories
  - Inverted pyramid structure
    - The first n sentences are often the best summary of length n
- Some genres are hard to summarize
  - Long documents (novels, the bible)
  - Scientific articles?
- Trainable summarizers are genre-specific.

Heuristic Method

- Luhn 58
  - http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5392672
- Edmundson
  - http://dl.acm.org/citation.cfm?doid=321510.321519

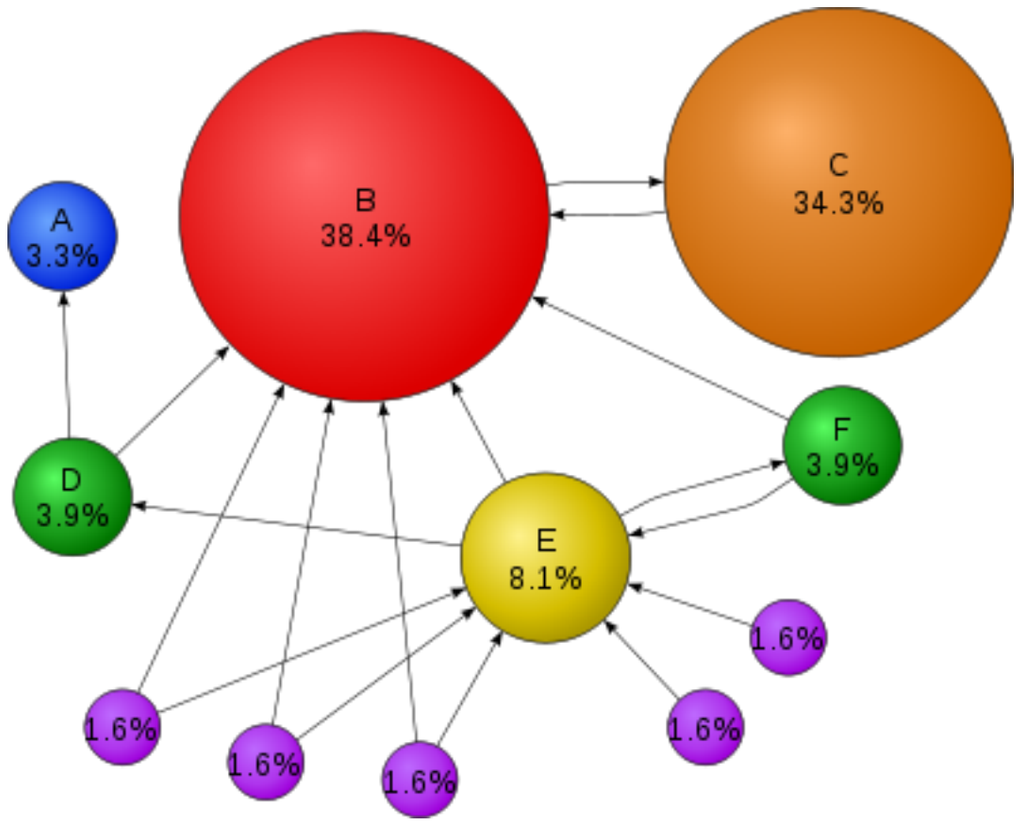
Supervised

- Naive Bayes
- Maximum Entropy
- SVM

- Feature can be
  - Sentence position, length, TF-IDF, …

PageRank

- A Ranking Algorithm
- Used by early Google Search



Unsupervised

- TextRank
- LexRank

- Both are graph-based algorithm that run PageRank
- Edge is undirected and determined by semantic similarity or content overlap

Unsupervised

- TextRank
  - Single document
- LexRank
  - Multi-document
  - Risk of selecting duplicate or highly redundant sentences
  - Require post-processing steps (Cross-Sentence Information Subsumption)

Supervised vs Unsupervised

- Supervised method require summaries manually created by extracting sentences
(not how human generate summarization)

- Very expensive to produce and relatively sparse

- Domain dependent and independent

Metrics

- ROUGE score
  - Recall-based
  - Unigram, Bigram, …
    - ROUGE-1, ROUGE-2, ROUGE-SU4
- Human evaluation
  - Pyramid
  - NIST

ROUGE Score

- Reference-summary: **Beijing** hosted **the summer Olympics**
- System-summary: **The summer Olympics** were held in **Beijing**
- ROUGE-1 score: 0.75

- Reference-summary: **The policemen killed the gunman**
- System-summary: **The gunman killed the policemen**
- ROUGE-1 score: 1

ROUGE Score

- The ROUGE score is averaged for multiple references.

- ROUGE-1 does not determine if the result is coherent or if the sentences flow together in a sensible manner.

- A higher order n-gram ROUGE score can measure fluency to some degree.

Data

- DUC2001 - 2011

  http://duc.nist.gov/data.html (2001-2007)

  https://tac.nist.gov/data/index.html (2008-2011)

- TREC Temporal Summarization Track (2013-2015)

  http://www.trec-ts.org/downloads

Baselines

- LEAD-3
- SumBasic
  - http://www.cis.upenn.edu/~nenkova/papers/ipm.pdf

System Demo

- Columbia Newsblaster

- http://newsblaster.cs.columbia.edu

Project Ideas

on Text Summarization

NBA auto highlights

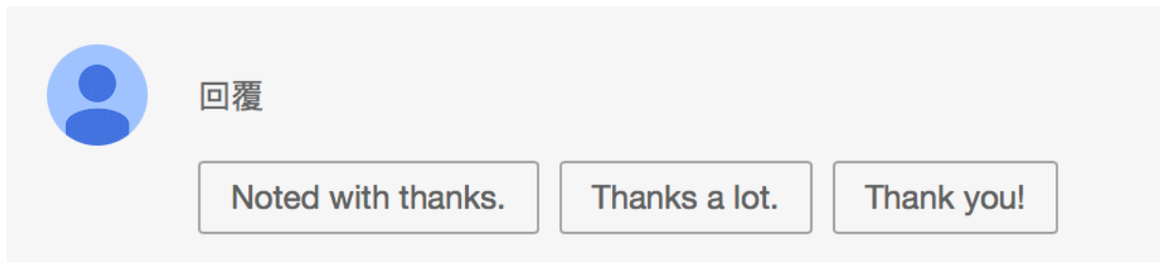
- 150 min full game into 3-4 min highlights
- Identify and extract highlight segments (extractive)
- Give description and voiceover (abstractive)

NBA auto highlights

- Play-by-play description and its video segment
- Player(s), action, time, score, …
- Ranking and Classification problem
- Need replay Slow motion or from different angles?

E-mail

- Email preview now: first-n words
- Identify the main content in the mail
- Response template (Sequence-to-Sequence Models)



Aided Summarization System

- A growing trend
- An HCI research topic
  - Machine help human generate sum
  - Human help machine generate sum
  - …
- Gather keyphrase, additional information retrieval, concept map generation, …

Inbox by Gmail

無需開啟郵件，就能直接查看訂單的最新情況、班機狀態、各項預約的詳細資料和郵件圖片。

今天

Amazon.com

您的 Amazon.com 訂單商品已出貨！

Urbanears 耳罩式耳機

預計送達日期：11 月 22 日

追蹤送貨進度

Virgin America

Virgin America 機位預訂資訊

Virgin America 班機號碼：12

SFO–JFK 11 月 15 日上午 6:55

辦理登機

Summarization

- A procedure of reducing text while retaining important points

- Extractive/abstractive

- Keyword vs. Graph-based

- Keyword based techniques rank sentences based on the occurrence of relevant keywords.

- Graph based techniques rank sentences based on content overlap.

Lab 9: play with sumy

- https://github.com/miso-belica/sumy