Summarization

Because time is money

Lecture 9: Text Analytics for Big Data

Terry Szymanski, Insight/CSI, UCD

Selling Things

stock-markets

social media

science

news

polls

sentiment-id

sentiment-use

time-series

summaries

VSMs

Classifiers

Clustering

*** cosine jaccard dice levenschtein ***

TF-IDF LLR PMI Entropy

simple frequencies

Outline

- Background and overview of summarization
- Basic methods for automatic summarization
  - Frequency-based extractive summarization
  - Python code example
- Extensions to the basic method
  - Improving summary outputs
  - Recent work and applications

Summarization

Background

(Why summarize anyway?)

What is summarization?

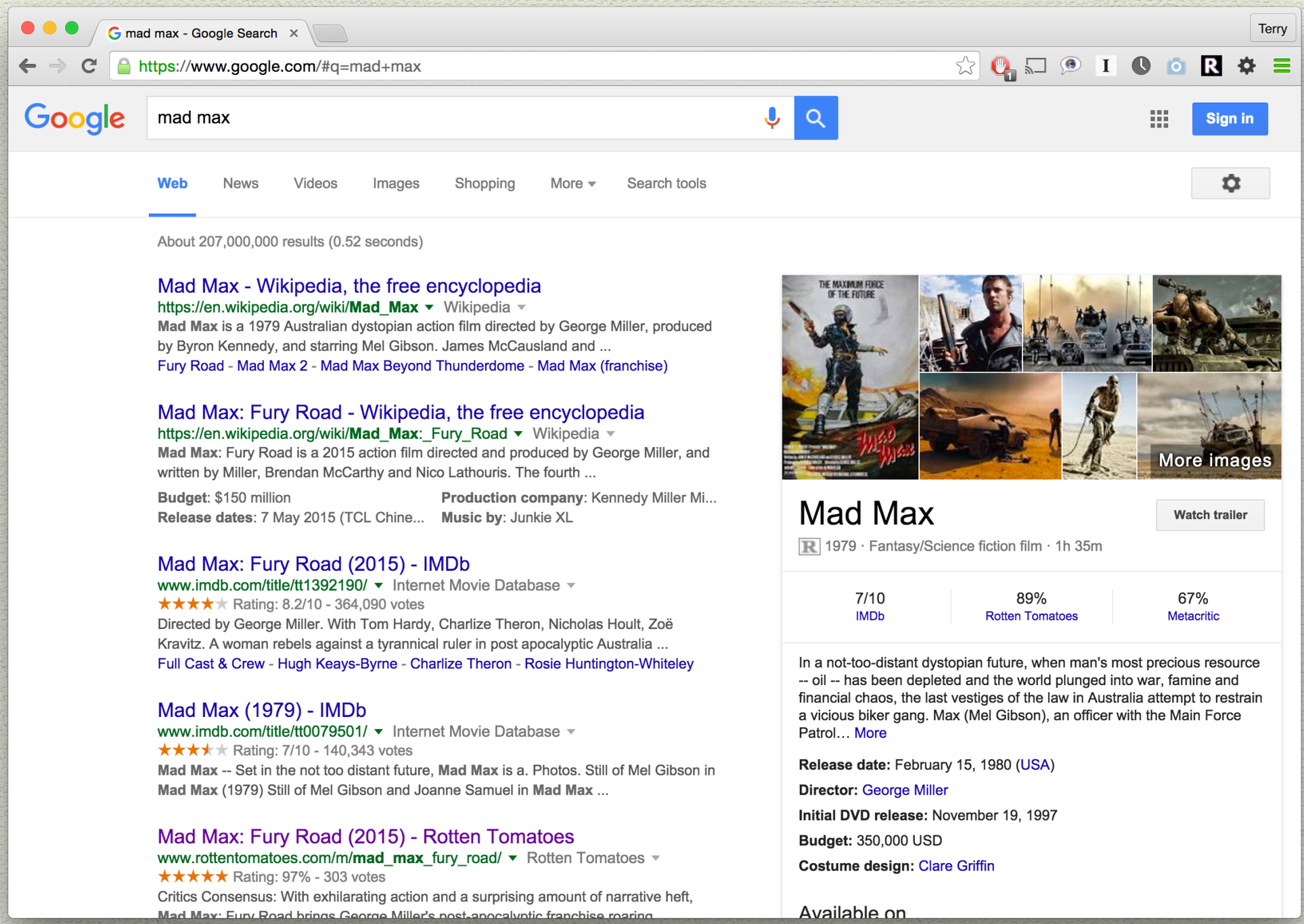
- "Text summarization is the process of distilling the most important information from a text to produce an abridged version for a particular task and user."

(Jurafsky & Martin 2009, adapted from Mani and Maybury 1999)

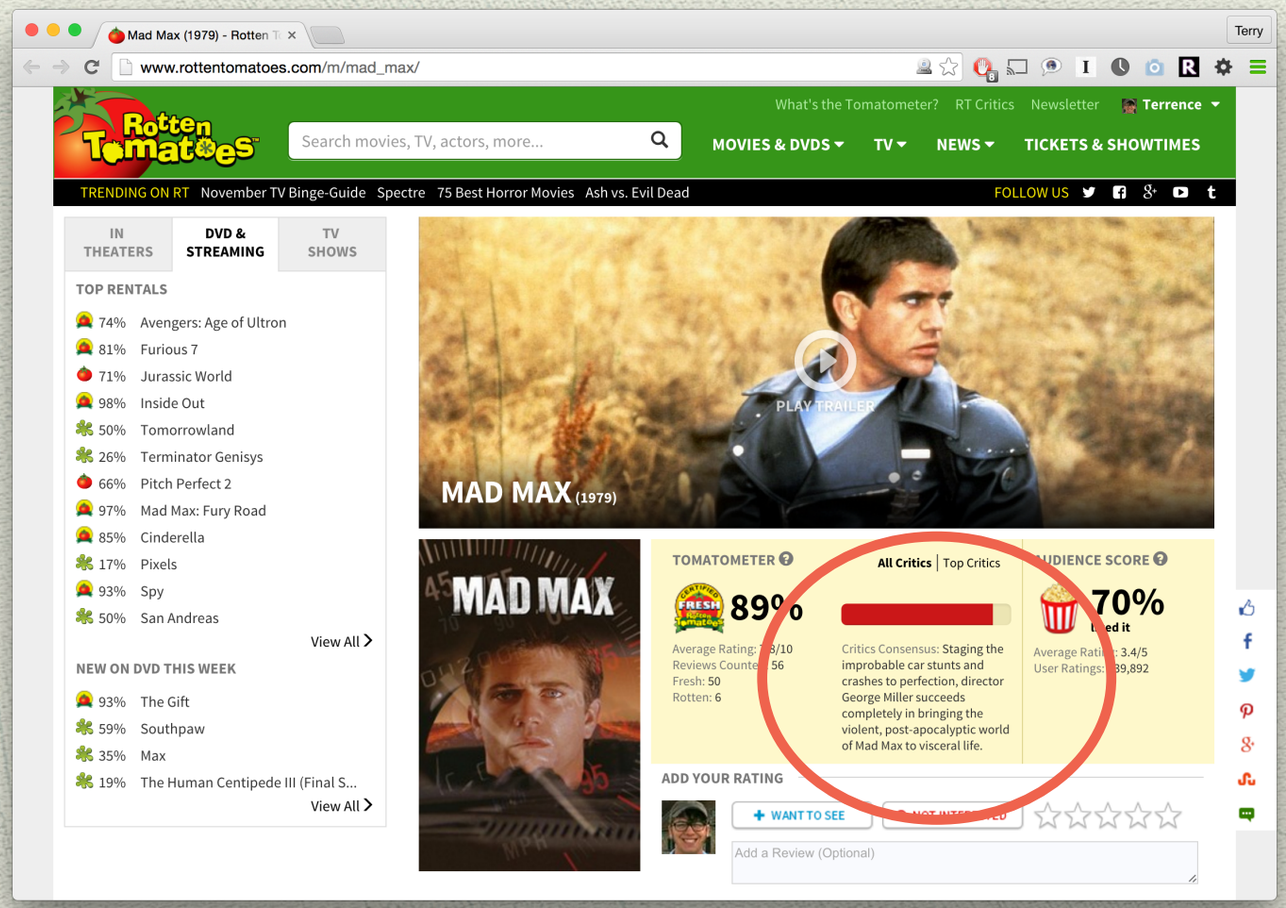
Examples of real-world summaries

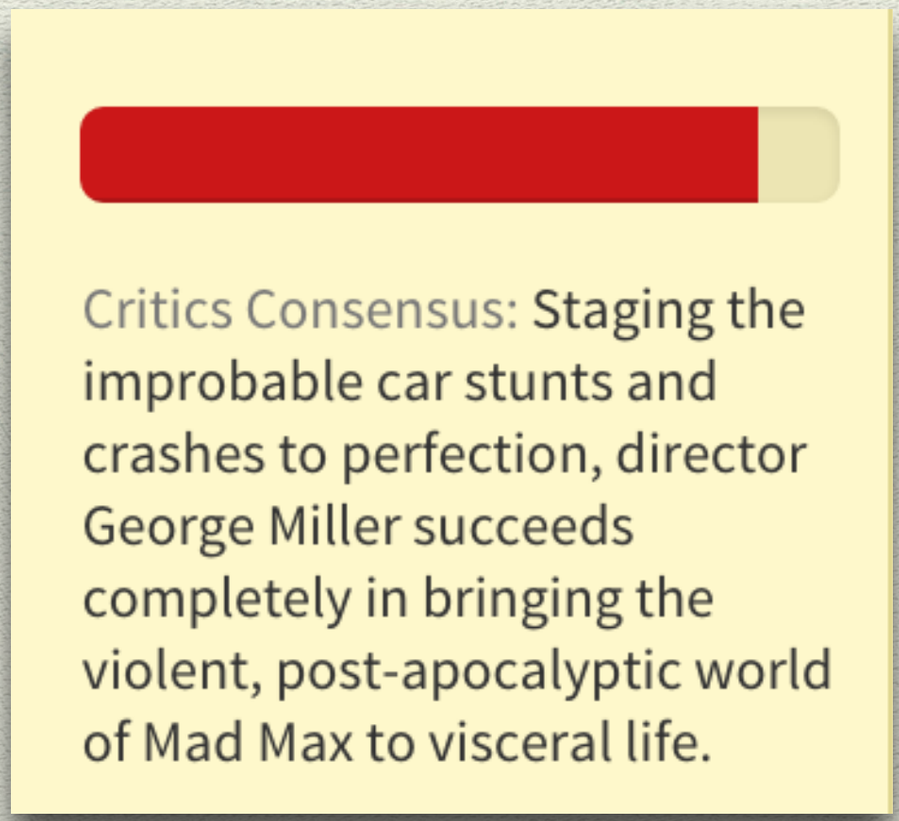
- Plot summaries
  - Literature (Cliff's / Spark / York Notes)
  - Film plot summaries (Wikipedia, IMDB, RT)
- Scientific paper abstracts
- Outlines
- Newspaper headlines?
- Titles?
- 5-Second Movies
- Google search / news snippets

Example: Search snippets



Example: Summary of reviews





Combine multiple reviews into a single sentence (manually)

Example: Plot Summary

mad max summary - Goog

Terry

https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=mad%20max%20su…

Google

mad max summary

Sign in

Web Videos News Images Shopping More Search tools

About 3,030,000 results (0.60 seconds)

Mad Max: Fury Road / Film synopsis

Years after the collapse of civilization, the tyrannical Immortan Joe enslaves apocalypse survivors inside the desert fortress the Citadel. When the warrior Imperator Furiosa (Charlize Theron) leads the despot's five wives in a daring escape, she forges an alliance with Max Rockatansky (Tom Hardy), a loner and former captive. Fortified in the massive, armored truck the War Rig, they try to outrun the ruthless warlord and his henchmen in a deadly high-speed chase through the Wasteland.

Feedback

Mad Max: Fury Road (2015) - Plot Summary - IMDb
www.imdb.com/title/tt1392190/plotsummary ▾ Internet Movie Database ▾
Mad Max: Fury Road (2015) on IMDb: An apocalyptic story set in the furthest reaches of our planet, in a stark desert landscape where humanity is broken, and ...

Mad Max (1979) - Plot Summary - IMDb
www.imdb.com/title/tt0079501/plotsummary ▾ Internet Movie Database ▾
Mad Max (1979) on IMDb: A vision of an apocalyptic future set in the wastelands of Australia. Total social decay is just around the corner in this spectacular ...

Mad Max: Fury Road

R 2015 · Fantasy/Science fiction film · 2 hours

Watch trailer

8.2/10 · IMDb
89% · Metacritic
97% · Rotten Tomatoes

Years after the collapse of civilization, the tyrannical Immortan Joe enslaves apocalypse survivors inside the desert fortress the Citadel. When the warrior Imperator Furiosa (Charlize Theron) leads the despot's five wives in a daring escape, she forges an alliance with Max Rockatansky (Tom Hardy), … More

Release date: May 15, 2015 (USA)

Director: George Miller

Initial DVD release: September 1, 2015 (USA)

Film series: Mad Max

Screenplay: George Miller, Brendan McCarthy, Nick Lathouris

Available on

YouTube	From $3.99
iTunes	From $3.99
Amazon Video	From $3.99

How to summarize?

- What is the source material?
  - Fiction? News? Technical?
  - Text? Audio? Video?
- What should the output look like?
  - 1 sentence? 3 sentences? 10? …
  - Bulleted list? Flowing prose?
- Who is the intended audience?
  - What do they already know?
- What is the intended purpose?
  - Guide people towards more info?
  - Eliminate need to read the full document?

There is no universal method for summarization

Why Summarize?

- Summaries take up **less time**
  - Help users to quickly decide what content to investigate further
  - May replace the need to read the original
- Summaries take up **less space**
  - Good for mobile devices
  - Many summaries can be on-screen at once

History of Automatic Summarization

- Goes way back:
  - Luhn (1958) extractive summarization using term frequency
  - Rath (1961) evaluation of auto- vs human extracts
  - Edmundson (1969) sentence weights, scientific paper extracts
- Early 2000s: news, multi-document summarization
  - DUC (2001-2006) shared datasets, evaluation
  - Columbia Newsblaster (2002)
- 2010s: speech / video, academic articles, biomedical…

Recent $ummarization Products

Wavii

- App to provide topic-based summaries (e.g. of people, sports teams, places)
- Bought by Google for $30m in 2013

Summly

- App to summarize news stories to display better on smartphones
- Bought by Yahoo! For $30m in 2013, making its creator a millionaire at age 15

Summarization

How to Summarize

(the basic method)

Extractive vs *Abstractive*

- **Extractive** summarization (→ *extracts*)
  - Build a summary by selecting phrases or sentences from the text
  - This is how most automatic systems work

- **Abstractive** summarization (→ *abstracts*)
  - Create a summary with novel wording that captures the main ideas of the document
  - This is usually how humans produce summaries
  - Very challenging to do automatically

Extractive Summarization

- Summary is formed by **extracting** snippets of text from the complete document.
- Basic method:

1. Assign a significance weight to each sentence in the document.
2. Select the *k* most-significant sentences to form a summary.

Abstractive Summarization

- Summary is formed by **abstracting** the main ideas from the original document
- Basic method:

1. Convert the text to a semantic (meaning) representation
2. Select components that belong in the summary
3. Generate a natural-language summary from the selected semantic content

Basic Extractive Summarization

- General extractive summarization method (Jurafsky & Martin 2009):
  1. **Content Selection**
     (choose a subset of sentences from document)
  2. **Information Ordering**
     (default: use order they appear in document)
  3. **Sentence Realization**
     (default: use sentences as found)

Example

- http://www.irishexaminer.com/breakingnews/sport/ireland-defeated-in-world-cup-quarter-final-against-argentina-701283.html

Example

HOME » SPORT

Ireland defeated in World Cup quarter-final against Argentina

7 2

Sunday, October 18, 2015 - 02:58 pm

Argentina piled on World Cup misery on Ireland once again by reaching the last four of the competition with a 43-20 victory in Cardiff this afternoon.

Ireland's record of never having reached the semi-final of a World Cup continued as Argentina scored four tries through Juan Imhoff (2), Matias Moroni and Joaquin Tuculet.

Ireland battled back from a first-quarter deficit of 17-0 with tries from Luke Fitzgerald and Jordi Murphy making it a three-point game at one stage, but the Pumas powered away in the closing stages to repeat their 1999 and 2007 World Cup victories over the Irish.

RELATED ARTICLES

Steve Hansen defends Richie McCaw over Francois Louw incident

South Africa v New Zealand: 5 lessons learned from the first World Cup semi-final

Nigel Owens will referee World Cup final, says All Blacks selector

Richie McCaw faces World Cup final ban if he is cited for elbowing Francois Louw

16 moments the New Zealand v South Africa Rugby World Cup semi-final could have gone either way

All Blacks through to Rugby World Cup final after tight win over South Africa

WATCH: Hundreds of supporters recreate famous Springboks run with Francois Pienaar

Schalk Burger brands All Blacks attack 'phenomenal'

Example

Natural 3-sentence summary

HOME » SPORT

Ireland defeated in World Cup quarter-final against Argentina

7 2

Sunday October 18, 2015 - 03:58 pm

Argentina piled on World Cup misery on Ireland once again by reaching the last four of the competition with a 43-20 victory in Cardiff this afternoon.

Ireland's record of never having reached the semi-final of a World Cup continued as Argentina scored four tries through Juan Imhoff (2), Matias Moroni and Joaquin Tuculet.

Ireland battled back from a first-quarter deficit of 17-0 with tries from Luke Fitzgerald and Jordi Murphy making it a three-point game at one stage, but the Pumas powered away in the closing stages to repeat their 1999 and 2007 World Cup victories over the Irish.

RELATED ARTICLES

- Steve Hansen defends Richie McCaw over Francois Louw incident
- South Africa v New Zealand: 5 lessons learned from the first World Cup semi-final
- Nigel Owens will referee World Cup final, says All Blacks selector
- Richie McCaw faces World Cup final ban if he is cited for elbowing Francois Louw
- 16 moments the New Zealand v South Africa Rugby World Cup semi-final could have gone either way
- All Blacks through to Rugby World Cup final after tight win over South Africa
- WATCH: Hundreds of supporters recreate famous Springboks run with Francois Pienaar
- Schalk Burger brands All Blacks attack 'phenomenal'

Lots of free summarizers

- http://www.summarizethis.com/
- http://textcompactor.com/
- http://autosummarizer.com/
- http://smmry.com/
- https://www.tools4noobs.com/summarize/

summarizethis.com

SummarizeThis™

AN **IRIS READING** PRODUCTIVITY TOOL | ABOUT IRIS READING | LIVE COURSES | ONLINE COURSES | FREE RESOURCES

Just the facts

Summary of content

Close

Argentina piled on World Cup misery on Ireland once again by reaching the last four of the competition with a 43-20 victory in Cardiff this afternoon. Ireland's record of never having reached the semi-final of a World Cup continued as Argentina scored four tries through Juan Imhoff (2), Matias Moroni and Joaquin Tuculet. Ireland battled back from a first-quarter deficit of 17-0 with tries from Luke Fitzgerald and Jordi Murphy making it a three-point game at one stage, but the Pumas powered away in the closing stages to repeat their 1999 and 2007 World Cup victories over the Irish. Ireland were featuring in their sixth World Cup quarter-final having not won any of the previous five while Argentina were chasing a second semi-final after making the last four in 2007. Argentina had not beaten Ireland since a pool game in that tournament, with the men in green winning the five matches between the two countries since. Ireland started as favourites but the absence of injured trio Sexton, O'Connell and O'Mahony as well as the suspended O'Brien gave Argentina belief that they would be able to bridge the gap. And they started in spectacular fashion with a third-minute try as full-back Joaquin Tuculet took a battle and flanker Pablo Matera made hard yards down the middle. The ball went quickly through the hands of Nicolas Sanchez, Juan Martin Fernandez Lobbe and Santiago Cordero for Bosch's replacement Matias Moroni to make the corner, Sanchez converting for a seven-point lead.

Argentina had not beaten Ireland since a pool game in that tournament, with the men in green winning the five matches between the two countries since.

Irelands injury curse struck again when Tommy Bowe left the field on a cart and Sanchez extended Argentinas lead to 17 points with a well-struck penalty.

However, Argentina restored their 17-point advantage when Ireland flanker Henry was penalised for entering the wrong side of a ruck and Sanchezs trusty boot did the rest.

Ireland were belatedly finding some rhythm and Madigans penalty hit the post - the third time an upright had been struck while the Pumas escaped when Sanchez had the ball ripped from his grasp in trying to relieve the pressure.

Ireland were applying pressure at the scrum as the contest headed towards its final quarter, but Madigan failed to take an opportunity to level the scores with a kick from just inside the Argentina half.

freesummarizer.com

Summarize Now!

Here is the summary:

But Argentina's momentum was slowed by the sin-binning of prop Ramiro Herrera for a late tackle on Keith Earls and Madigan finally put Ireland points on the scoreboard with a straightforward penalty.

Ireland were belatedly finding some rhythm and Madigan's penalty hit the post - the third time an upright had been struck while the Pumas escaped when Sanchez had the ball ripped from his grasp in trying to relieve the pressure.

But Argentina's momentum was slowed by the sin-binning of prop Ramiro Herrera for a late tackle on Keith Earls and Madigan finally put Ireland points on the scoreboard with a straightforward penalty.

tools4noobs.com

Summary:

1. Ireland battled back from a first-quarter deficit of 17-0 with tries from Luke Fitzgerald and Jordi Murphy making it a three-point game at one stage, but the Pumas powered away in the closing stages to repeat their 1999 and 2007 World Cup victories over the Irish.
2. Ireland were belatedly finding some rhythm and Madigan's penalty hit the post - the third time an upright had been struck – while the Pumas escaped when Sanchez had the ball ripped from his grasp in trying to relieve the pressure.
3. Ireland were featuring in their sixth World Cup quarter-final having not won any of the previous five while Argentina were chasing a second semi-final after making the last four in 2007.

smmry.com

SMMRY

This is a 3 sentence summary of the text you submitted. Public SAVE

Summary processing at low priority, upgrade to BOOST

Ireland's record of never having reached the semi-final of a World Cup continued as Argentina scored four tries through Juan Imhoff, Matias Moroni and Joaquin Tuculet.

Ireland's injury curse struck again when Tommy Bowe left the field on a cart and Sanchez extended Argentina's lead to 17 points with a well-struck penalty.

Argentina restored their 17-point advantage when Ireland flanker Henry was penalised for entering the wrong side of a ruck and Sanchez's trusty boot did the rest.

Text style: Expanded | Compact Reduction: 91 % Characters: 481 Fixed height: Enabled | Disabled

SETTINGS SHARE SUMMARY RATE SUMMARY NEW SUMMARY

Baseline: First-*k* Sentences

- Choose the first *k* sentences of the document to serve as the summary

- Very simple, but hard to beat, especially for very short summaries (e.g. headlines)
  - Most news articles are written in the "inverted pyramid" structure: the most important pieces of information are given first.

- Alternative baseline: choose *k* random sentences

Sentence Scoring

Extractive summarization generally involves

1. scoring sentences
2. choosing sentences with high scores

- There are multiple ways to calculate these scores.

Word-Weight Methods

- Bag-of-words: sentence weight is sum of word weights
- Weights can be calculated in different ways (frequency, tf-idf, likelihood ratio, …)
- This can be viewed as finding sentences that are closest to an imaginary "centroid" sentence

1. Assign each word a weight based on how "salient" it is
2. Score each sentence by summing its constituent words' weights
3. Choose the highest-scored sentences

SumBasic

- SumBasic – count the number of "frequent" words in the sentence S

- p(w) = relative frequency of word w in the document.

$$score(S) = \sum_{w \in S} \frac{1}{|S|} p_D(w)$$

TF-IDF

- Same idea, but use TF-IDF instead of simple frequency.

$$score(S) = \sum_{w \in S} \frac{1}{|S|} tfidf(w)$$

Recall: Log-Likelihood Ratio

$$LLR = -2 \log \lambda$$

$$\lambda = \frac{L(H_0)}{L(H_1)}$$

$$H_0 : \quad \theta = \theta_0,$$
$$H_1 : \quad \theta = \theta_1.$$

$$L(n, k, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

LLR is minus 2 times log of lambda

Lambda is the likelihood of null hypothesis (H_0) over alternative (H_1)

L is computed relative to the binomial distribution

Topic Words (LLR)

1. Calculate log-likelihood ratio λ of all words in document.
   - (Is this word significantly more likely to appear in *this* document as opposed to a reference corpus of documents?)
2. Identify "topic words" as terms with a minimum λ score.

- Unlike other metrics, the weights are binary (1 or 0)
- These types of features tend to outperform frequency features on certain tasks (e.g. multi-document news summarization)

$$t(w) = \begin{cases} 1 & if \ -2\log\lambda(w) > 10 \\ 0 & otherwise \end{cases}$$

$$score(S) = \sum_{w \in S} \frac{1}{|S|} t(w)$$

Frequency Example

Input Document

Argentina piled on World Cup misery on Ireland once again by reaching the last four of the competition with a 43-20 victory in Cardiff this afternoon.

Ireland's record of never having reached the semi-final of a World Cup continued as Argentina scored four tries through Juan Imhoff (2), Matias Moroni and Joaquin Tuculet.

Ireland battled back from a first-quarter deficit of 17-0 with tries from Luke Fitzgerald and Jordi Murphy making it a three-point game at one stage, but the Pumas powered away in the closing stages to repeat their 1999 and 2007 World Cup victories over the Irish.

Ireland were featuring in their sixth World Cup quarter-final having not won any of the previous five while Argentina were chasing a second semi-final after making the last four in 2007.

Argentina had not beaten Ireland since a pool game in that tournament, with the men in green winning the five matches between the two countries since.

....

1. Tokenize text
2. Lowercase
3. Remove stop words
4. Count token frequencies

Frequency Example

Input Document

Argentina piled on World Cup misery on Ireland once again by reaching the last four of the competition with a 43-20 victory in Cardiff this afternoon.

Ireland's record of never having reached the semi-final of a World Cup continued as Argentina scored four tries through Juan Imhoff (2), Matias Moroni and Joaquin Tuculet.

Ireland battled back from a first-quarter deficit of 17-0 with tries from Luke Fitzgerald and Jordi Murphy making it a three-point game at one stage, but the Pumas powered away in the closing stages to repeat their 1999 and 2007 World Cup victories over the Irish.

Ireland were featuring in their sixth World Cup quarter-final having not won any of the previous five while Argentina were chasing a second semi-final after making the last four in 2007.

Argentina had not beaten Ireland since a pool game in that tournament, with the men in green winning the five matches between the two countries since.

….

term	frequency
ireland	17
argentina	16
sanchez	10
game	7
ball	6
final	6
penalty	6
madigan	6
fitzgerald	5
half	5
four	4
try	4
making	4
imhoff	4
irish	4
pumas	4
cup	4
world	4

Frequency Example

Input Document

Argentina piled on World Cup misery on Ireland once again by reaching the last four of the competition with a 43-20 victory in Cardiff this afternoon.

Ireland's record of never having reached the semi-final of a World Cup continued as Argentina scored four tries through Juan Imhoff (2), Matias Moroni and Joaquin Tuculet.

Ireland battled back from a first-quarter deficit of 17-0 with tries from Luke Fitzgerald and Jordi Murphy making it a three-point game at one stage, but the Pumas powered away in the closing stages to repeat their 1999 and 2007 World Cup victories over the Irish.

Ireland were featuring in their sixth World Cup quarter-final having not won any of the previous five while Argentina were chasing a second semi-final after making the last four in 2007.

Argentina had not beaten Ireland since a pool game in that tournament, with the men in green winning the five matches between the two countries since.

….

1. Split document into sentences
2. Each sentence is a list of tokens

Frequency Example

Input Document

Argentina piled on World Cup misery on Ireland once again by reaching the last four of the competition with a 43-20 victory in Cardiff this afternoon.

Ireland's record of never having reached the semi-final of a World Cup continued as Argentina scored four tries through Juan Imhoff (2), Matias Moroni and Joaquin Tuculet.

Ireland battled back from a first-quarter deficit of 17-0 with tries from Luke Fitzgerald and Jordi Murphy making it a three-point game at one stage, but the Pumas powered away in the closing stages to repeat their 1999 and 2007 World Cup victories over the Irish.

Ireland were featuring in their sixth World Cup quarter-final having not won any of the previous five while Argentina were chasing a second semi-final after making the last four in 2007.

Argentina had not beaten Ireland since a pool game in that tournament, with the men in green winning the five matches between the two countries since.

....

```
0: argentina piled on world cup misery on
ireland once again by reaching the last
four of the competition with a 43 - 20
victory in cardiff this afternoon .

1: ireland ' s record of never having
reached the semi - final of a world cup
continued as argentina scored four tries
through juan imhoff ( 2 ), matias moroni
and joaquin tuculet .

2: ireland battled back from a first -
quarter deficit of 17 - 0 with tries from
luke fitzgerald and jordi murphy making it
a three - point game at one stage , but
the pumas powered away in the closing
stages to repeat their 1999 and 2007 world
cup victories over the irish .

...
```

Frequency Example

Calculating score of sentence 1

```
argentina piled on world cup misery on
ireland once again by reaching the last
four of the competition with a 43 - 20
victory in cardiff this afternoon .
```

Score = 16 + 1 + 0 + 4 + 4 + …

Score = 55

Word	Freq
argentina	16
piled	1
on	0
world	4
cup	4
misery	1
on	0
ireland	17
once	0
again	0
by	0
reaching	1
the	0
last	2
four	4
of	0
the	0
competition	1
with	0
a	0
43	0
-	0

Frequency Example

- Score each sentence:

Score	Index	Tokens
55	0	argentina piled on world cup misery on ireland once again by reaching the last f
78	1	ireland ’ s record of never having reached the semi - final of a world cup conti
83	2	ireland battled back from a first - quarter deficit of 17 - 0 with tries from lu
78	3	ireland were featuring in their sixth world cup quarter - final having not won a
56	4	argentina had not beaten ireland since a pool game in that tournament , with the
64	5	ireland started as favourites but the absence of injured trio sexton , o ’ conne
36	6	and they started in spectacular fashion with a third - minute try as full - back
59	7	the ball went quickly through the hands of nicolas sanchez , juan martin fernand
43	8	ireland were shell - shocked but worse was to come after 10 minutes when another
46	9	it was touch and go whether wing juan imhoff would reach the ball in time to gro
75	10	ireland ’ s injury curse struck again when tommy bowe left the field on a cart a

Frequency Example

- Sort by score:

Score	Index	Tokens
83	2	ireland battled back from a first - quarter deficit of 17 - 0 with tries from lu
78	1	ireland ' s record of never having reached the semi - final of a world cup conti
78	3	ireland were featuring in their sixth world cup quarter - final having not won a
75	10	ireland ' s injury curse struck again when tommy bowe left the field on a cart a
75	14	ireland were belatedly finding some rhythm and madigan ' s penalty hit the post
72	21	ireland were applying pressure at the scrum as the contest headed towards its fi
66	28	the outside - half took his final total to 23 with his fourth conversion and fif
66	22	but argentina ' s momentum was slowed by the sin - binning of prop ramiro herrer
64	12	sanchez struck the post with another long - range attempt and ireland made the m
64	5	ireland started as favourites but the absence of injured trio sexton , o ' conne
64	11	however , argentina restored their 17 - point advantage when ireland flanker hen

Frequency Example

- Generate a summary by choosing the top-*k* sentences:

  - Ireland battled back from a first-quarter deficit of 17-0 with tries from Luke Fitzgerald and Jordi Murphy making it a three-point game at one stage, but the Pumas powered away in the closing stages to repeat their 1999 and 2007 World Cup victories over the Irish.
  - Ireland's record of never having reached the semi-final of a World Cup continued as Argentina scored four tries through Juan Imhoff (2), Matias Moroni and Joaquin Tuculet.
  - Ireland were featuring in their sixth World Cup quarter-final having not won any of the previous five while Argentina were chasing a second semi-final after making the last four in 2007.

Live Code Example

- Let's see this example in more detail.
- (switch to iPython Notebook)

Graph Methods

- Intuition: graph-based concept of importance
  - Google PageRank: websites with lots of inbound links are important
  - Social networks: people with lots of connections are important
  - Summarization: sentences with high centrality are important

Graph-based Methods

- *LexRank* (Erkan and Radev 2004)
- Each sentence is a node in a graph
- Edges are similarity between sentences (e.g. cosine similarity)

1. Calculate a distance/similarity score between each pair of sentences.
2. Score each sentence based on its "centrality" (how well-connected it is to other sentences)
3. Choose the highest-scored sentences

LexRank

Multi-document input

SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq "did not end" and that Britain is still "ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq "will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

Calculate similarity between pairs of sentences

	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

Construct a graph and find the most central nodes

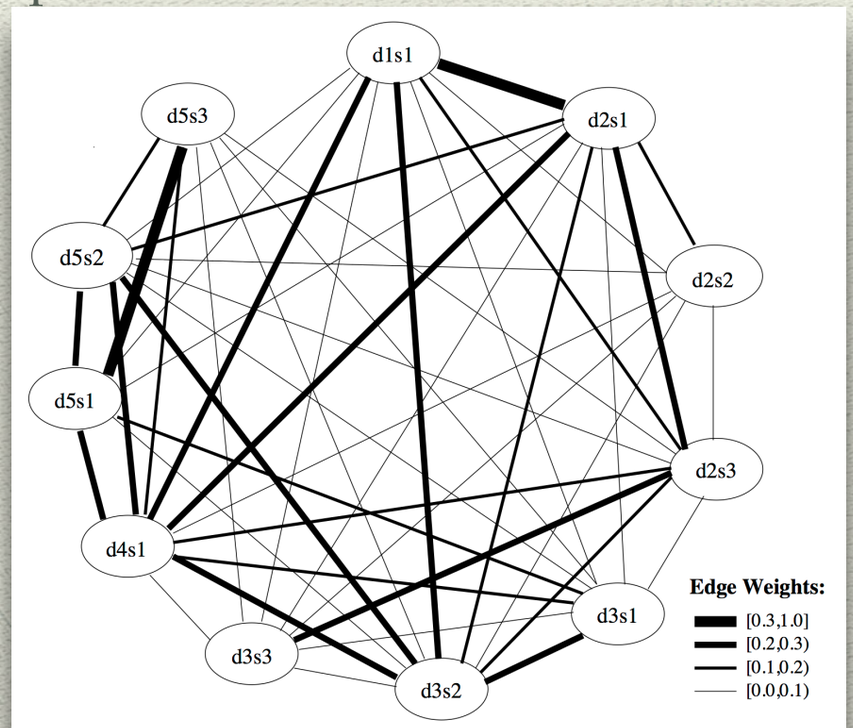


Figure 2: Weighted cosine similarity graph for the cluster in Figure 1.

Supervised Classification

- Sentence extraction can be viewed as binary classification

1. Create a labeled data set
   - each sentence in a document is either (1) in the summary or (0) not in the summary
2. Train a binary classifier
   - Use any features: weights, position in doc, specific terms
   - Naive Bayes, SVM, decision trees…
3. Select the (1)-labeled sentences

Supervised Classification Methods

- PRO: Can use a variety of features
- CON: Getting labeled data is hard
- CON: Binary classification may be too strict
  - Want to control how many sentences end up in summary
  - (better to try ranking or regression, e.g.)

Summarization

Evaluation

Evaluation Options

1. Compare candidate summary to one or more known good summaries as reference.
2. Directly assess the quality of the summary.
3. Indirectly measure the quality of the summary by using it in a related task.

- Evaluation is hard.
- Perceived quality depends on context and audience.
- Different wording may convey the same information.

Evaluation of Summaries

- **Manual evaluation**
  - Costly (time and effort) to compute
  - Can provide deeper analysis, and are closer to the actual objective (producing summaries that are good for humans)

- **Automatic evaluation (e.g. ROUGE)**
  - Easy to compute, and (in theory) compare across systems
  - Typically only capture shallow (e.g. n-gram) information

Pyramid Evaluation (Harnly 2005)

- Compare the semantic content of a candidate summary to a set of reference summaries.

1. Identify equivalent semantic content units (SCUs)
2. Count how many reference summaries contain each unit
3. Score candidate summary by adding the counts of its SCUs

- The counts have a "Pyramid" shape:
  - Small number of very-frequent units (the peak)
  - Large number of infrequent units (the base)
- The optimal summary contains the very-frequent units

SCU Examples

- Semantic Content Units (SCUs) are spans of text which express the same idea with possibly different wording.
- SCUs are *manually* identified and aligned across summaries for evaluation.

1. The cause of the Sept. 2, 1998 crash **has not been determined.**
2. Investigators of a Swissair crash that killed 229 people off the coast of Nova Scotia searched for clues as to a cause **but refrained from naming one.**
3. **The specific cause of the tragedy was never determined,** but suspicions are that an electrical short caused a fire.
4. Wreckage showed evidence of high heat and heat damaged wiring above the cockpit area but investigators remain **unsure of its cause.**

(example from Harnly et al. 2005)

Pyramid Evaluation

- Each SCU gets a weight w_{SCU} equal to the number of reference summaries that SCU appears in.
- The weight of a summary *S* is the sum weight of the SCUs in the summary.
- The score of *S is* the ratio of its weight to the optimal weight of a summary S* with an equal number of SCUs.

$$score(S) = \frac{\sum_{SCU \in S} w_{SCU}}{\sum_{SCU \in S^*} w_{SCU}}$$

Linguistic Quality Assessment

- Intuition: A good summary should be well-written and easy to read.
- Manually evaluate summaries on a five-point scale
- May be automated (Pitler et al. 2010)
  - Mimics human *overall ranking* of summarization *systems*

(1) Very Poor (2) Poor (3) Barely Acceptable (4) Good (5) Very Good

- **Grammaticality**: no errors that make text hard to read
- **Non-redundancy**: no unnecessary repetition of facts, names
- **Referential clarity**: easy to identify who pronouns refer to, etc.
- **Focus**: only contain related information
- **Structure and Coherence**: not just a heap of related information

http://www-nlpir.nist.gov/projects/duc/duc2006/quality-questions.txt

Intrinsic v Extrinsic Evaluation

- **Intrinsic** evaluation: measure the summary quality directly. E.g.
  - ROUGE, Pyramid method, quality assessment
- **Extrinsic** evaluation: measure how the summary affects performance on a different task. E.g.
  - question answering / reading comprehension
  - relevance assessment

Reading Comprehension

1. Ask people to read either a) the summary or b) the full text
2. Same people answer questions based on the content of the full text
3. Compare the results of group (a) and group (b) to see how well the summary captures the information of the full text

1.	Why does Trevor leave New York and where does he move to?
2.	What is KOS, who is their leader, and why is he attending high school?
3.	What happened to Cesar's finger, how did he eventually die?
4.	Who killed Benny and how does Ellen find out?
5.	Who is Rita and what becomes of her?

Table 1: Questions for the movie "One Eight Seven".

Example from
Gorinski and Lapata (2015)

Automatic Evaluation: ROUGE

ROUGE-N

$$= \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

- Measure which components (e.g. n-grams) of the candidate summary appear in (any of) the reference summaries (recall-based)
- ROUGE-2 is good for single-document summaries
- ROUGE-1 better for short summaries (e.g. headlines)

Problems with ROUGE

- Reliance on n-grams penalizes non-identical semantically equivalent units (e.g. synonyms).
- Many different flavors and parameters of ROUGE make it hard to compare across systems / papers.
- Modern systems often achieve similar ROUGE scores despite producing very different types of summaries (Hong 2014).

Summarization

Extensions and Improvements

(more ways to summarize)

Basic Extractive Summarization

- General extractive summarization method:
  1. **Content Selection**
     (choose a subset of sentences from document)
  2. **Information Ordering**
     (default: use order they appear in document)
  3. **Sentence Realization**
     (default: use sentences as found)

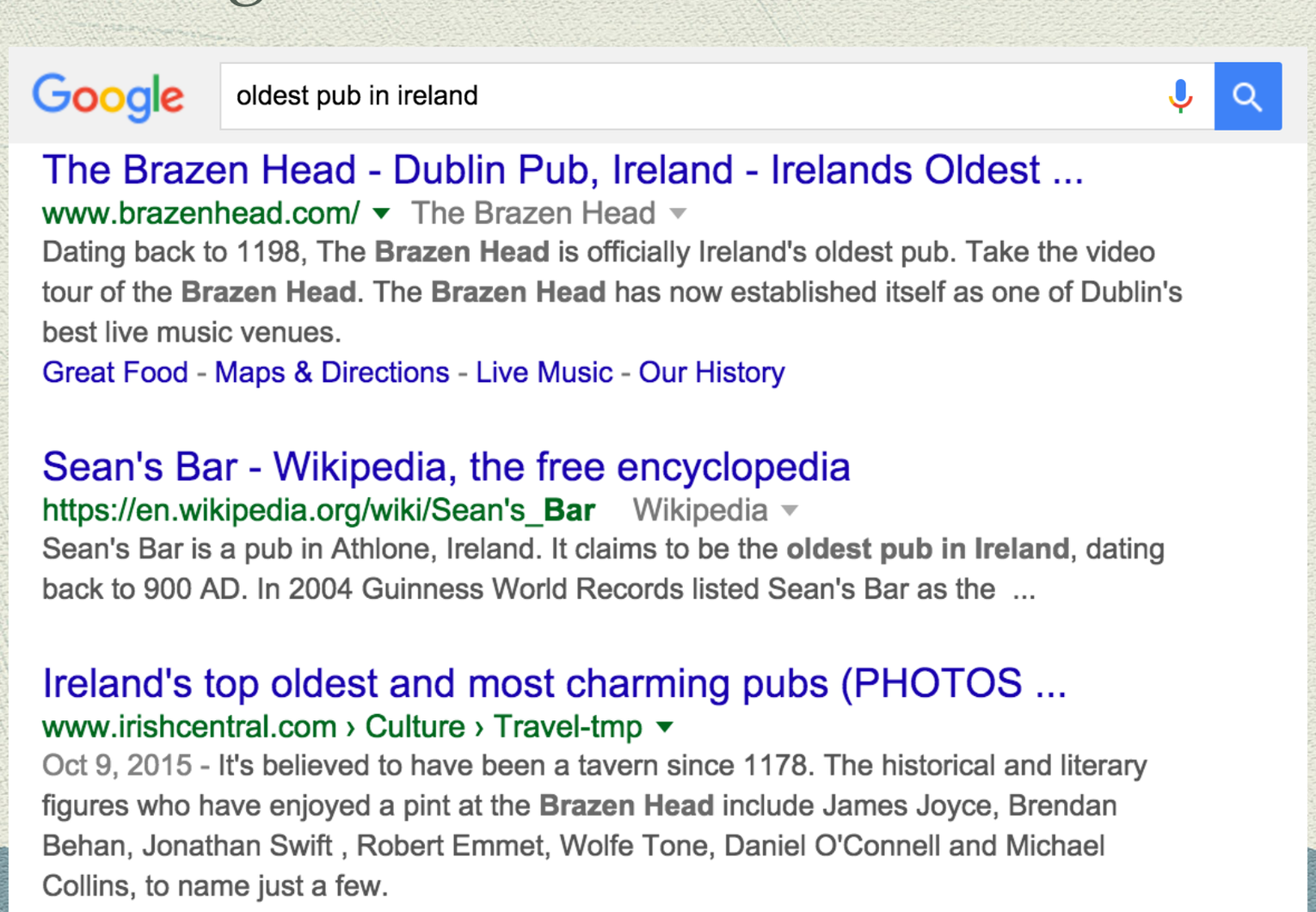
Multi-Document Summarization

Generate a summary for a *cluster* of related documents (e.g. news articles on the same story)

- Examples: search results, news clusters
- Ordering of sentences is more of a challenge than single-document clustering
- Need to avoid repeating the same information

Query-based summarization

- Can use extractive methods, increasing the weight of query terms
- Example: Google search results



Maximal Marginal Relevance

- Maximal relevance:
  - choose the sentence that contributes the most information about the document.
- Maximal *Marginal* Relevance:
  - choose the sentence that contributes the most information about the text *not already contained in the summary*. (Carbonell et al. 1998)

$$MMR = \arg\max_{D_i \in R-S}[\lambda Sim(D_i, Q) - (1-\lambda)\max_{D_j \in S} Sim(D_i, D_j)]$$

Theme Clustering

- Want to avoid selecting redundant sentences
  - Especially for multi-document clustering

1. Cluster sentences based on similarity into themes
2. Identify significant themes (ones containing most sentences)
3. Extract a representative sentence from each theme.

Information Ordering

How to order the sentences in an extractive summary?

1. Use ordering found in document (default)
   - Works well due to inherent document structure
   - Harder to use for multi-document summary
2. Order sentences based on chronological ordering of events
3. Order sentences based on weights / metric

Sentence Realization

Extractive summaries can be a bit “choppy”. Steps can be taken to clean up the output:

- Sentence simplification
  - Use syntactic structure to identify bits to cut out
- Coreference resolution
  - Clean up instances of “he”, “there” etc.
- Sentence combination
  - Merge multiple sentences into one

Parse Tree Pruning

- Dorr et al 2003. *Hedge Trimmer*
- Generate headline-like summaries by trimming sentences
  1. parse sentence
  2. prune parse tree according to various rules.

(1) Story Words: **Kurdish guerilla forces** *moving with lightning speed* **poured into Kirkuk** *today immediately* **after Iraqi troops**, *fleeing relentless U.S. airstrikes*, **abandoned** *the hub of Iraq's rich northern* **oil fields**.

Generated Headline: Kurdish guerilla forces poured into Kirkuk after Iraqi troops abandoned oil fields.

(12) Input: According to a now finalized blueprint described by U.S. officials and other sources, the Bush administration plans to take complete, unilateral control of a post-Saddam Hussein Iraq

Parse: **[S** *[PP According to a now-finalized blueprint described by U.S. officials and other sources] [Det the]* **Bush administration plans to take complete, unilateral control of** *[Det a]* **post-Saddam Hussein Iraq]**

Output of Preposed Adjunct Removal: Bush administration plans to take complete unilateral control of post-Saddam Hussein Iraq

Combining Multiple Sentences

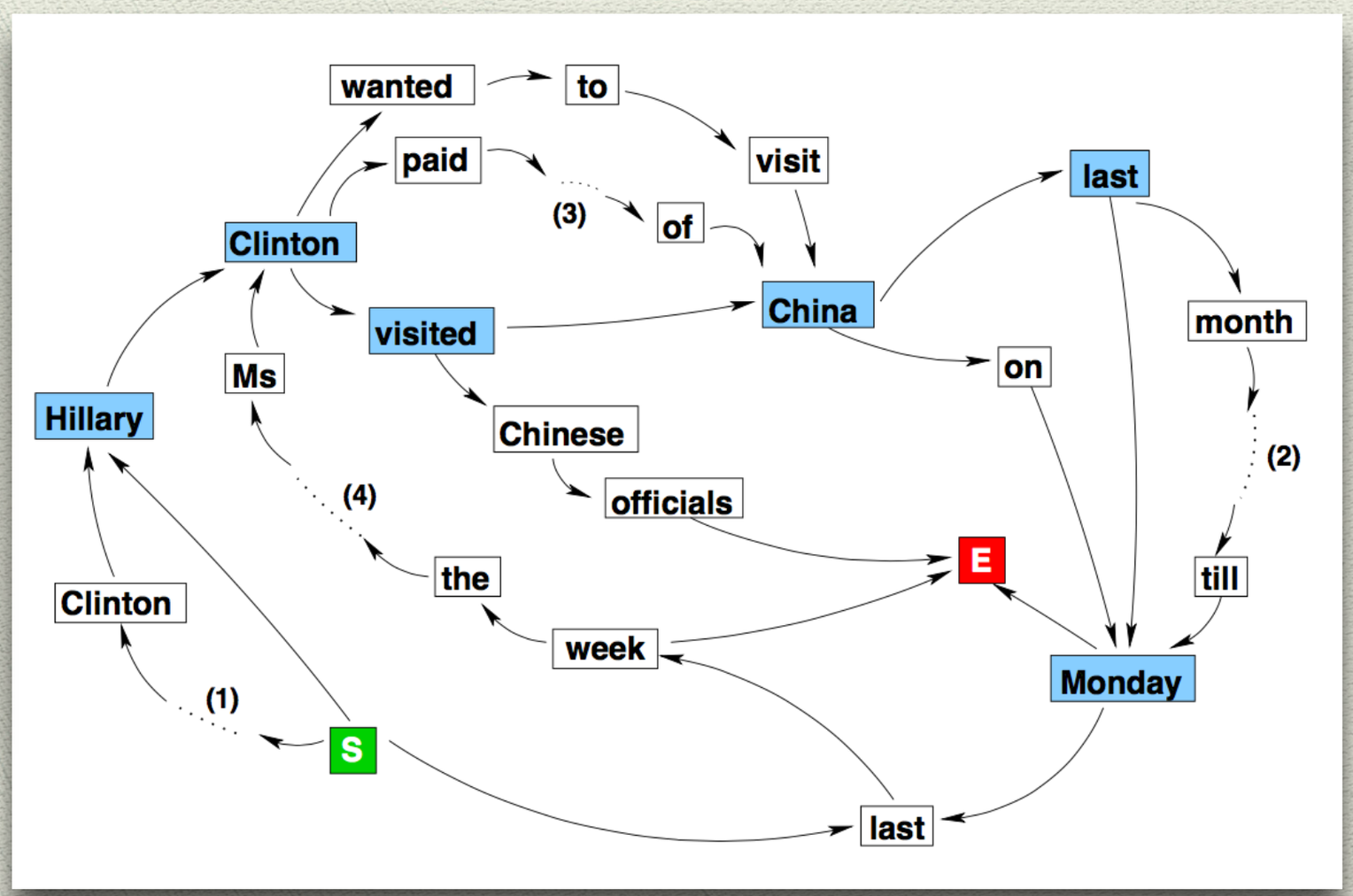
- Fillippova 2010:
  1. combine sentences into a directed graph
  2. find optimal path

(1) *The wife of a former U.S. president Bill* Clinton Hillary Clinton visited China last Monday.

(2) Hillary Clinton wanted to visit China last month *but postponed her plans* till Monday last week.

(3) Hillary Clinton paid *a visit to the People Republic* of China on Monday.

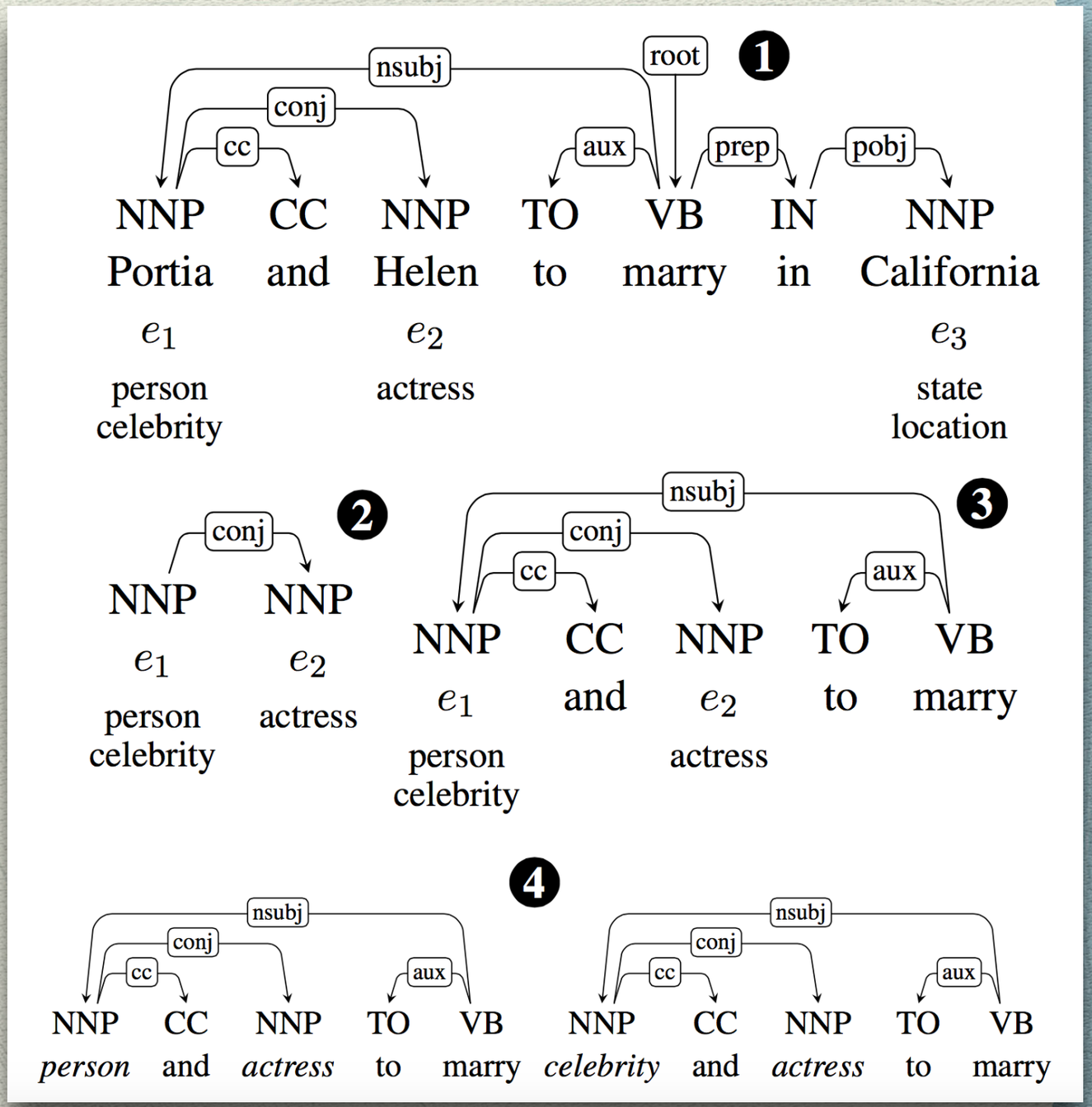
(4) Last week the *Secretary of State* Ms. Clinton visited Chinese officials.



Learning headline templates

- Alfonseca et al. 2013. *HEADY*
  1. Parse headlines from a news cluster
  2. Identify named entities
  3. Infer headline templates

- Carmelo and La La Party It Up with Kim and Ciara
- La La Vazquez and Carmelo Anthony: Wedding Day Bliss
- Carmelo Anthony, actress LaLa Vazquez wed in NYC
- Stylist to the Stars
- LaLa, Carmelo Set Off Celebrity Wedding Weekend
- Ciara rocks a sexy Versace Spring 2010 mini to LaLa Vasquez and Carmelo Anthony's wedding (photos)
- Lala Vasquez on her wedding dress, cake, reality tv show and fiancé, Carmelo Anthony (video)
- VAZQUEZ MARRIES SPORTS STAR ANTHONY
- Lebron Returns To NYC For Carmelo's Wedding
- Carmelo Anthony's stylist dishes on the wedding
- Paul pitching another Big Three with "Melo in NYC"
- Carmelo Anthony and La La Vazquez Get Married at Star-Studded Wedding Ceremony



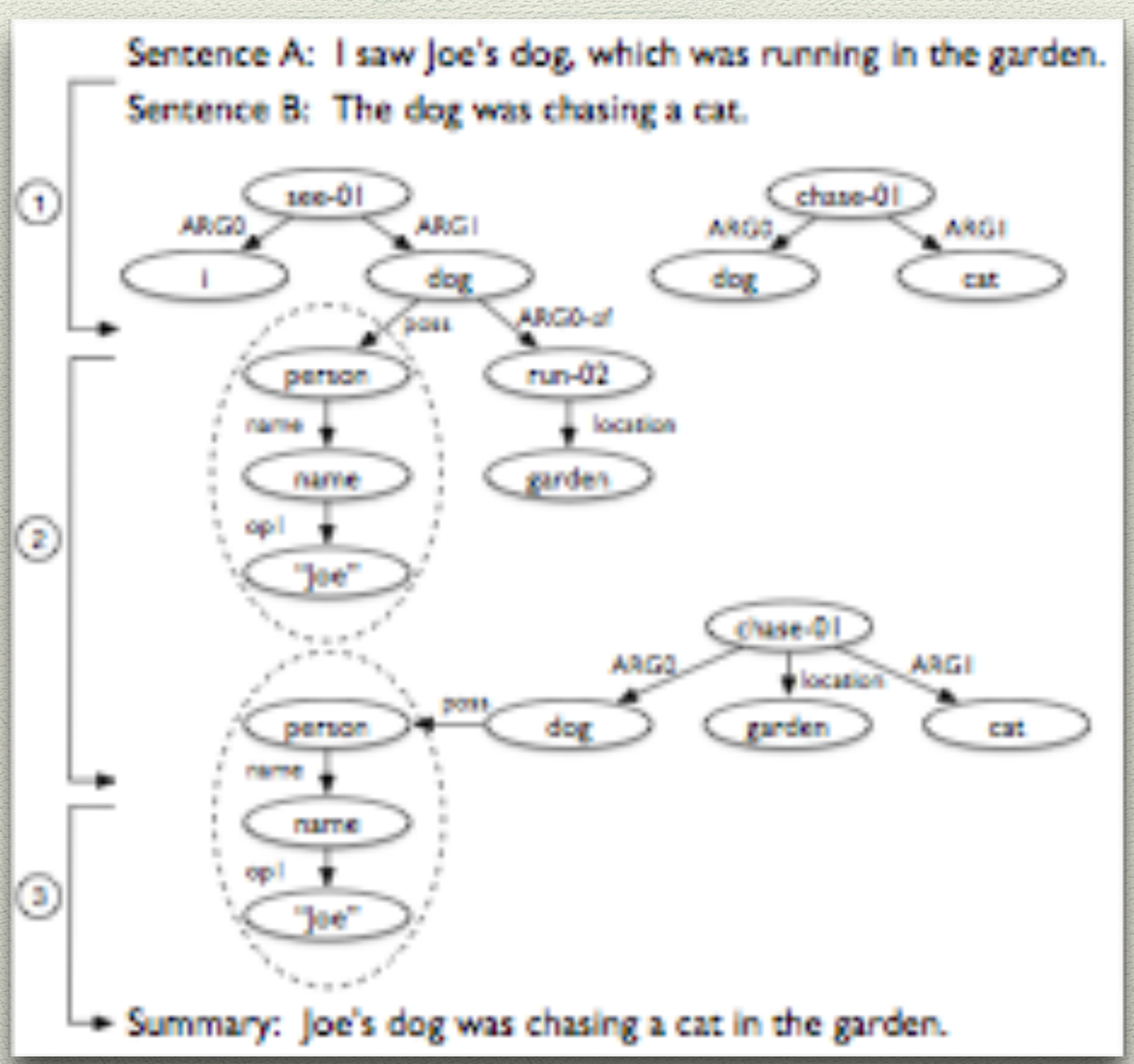
Abstractive Summarization

Liu et al. (2015)

1. Parse sentences to AMR (Abstract Meaning Representation)
2. Transform AMR graphs into a single summary graph
3. Generate English summary from the summary AMR graph

Challenges:

- Each step (1-3) is noisy
- No AMR text-generation system currently exists



Movie Script Summarization

- Gorinski and Lapata. 2015
- Extract *scenes* from a movie script to produce a summary for film executives to read
- Extract *chains* of scenes using various criteria
- Evaluate using reading comprehension measures

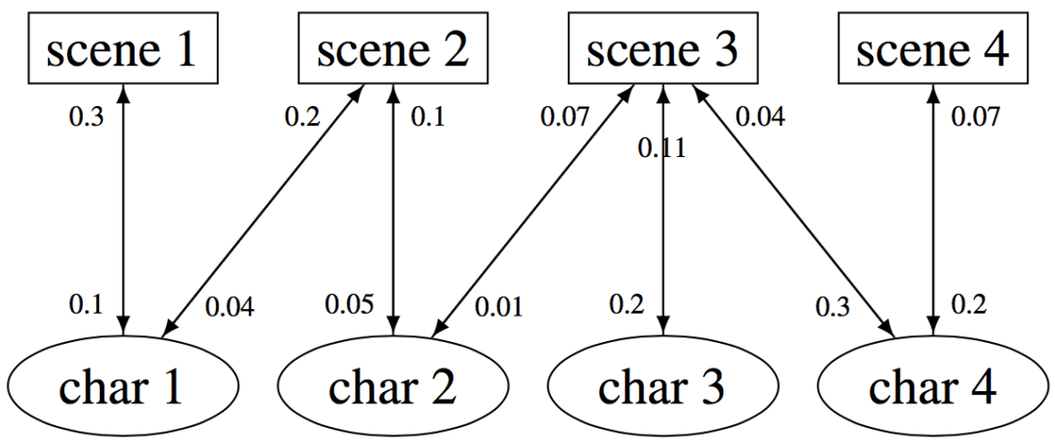


Figure 4: Example of a bipartite graph, connecting a movie's scenes with participating characters.

Recent Trends in Summarization

New domains for summarization:

- Academic literature
  - View citations as summaries
  - Automatic creation of literature reviews
- Microblogs / twitter
- Video summarization

Challenges for current systems (Nenkova?):

- How to generate “smooth” flowing text?
- Summarization of speech (and video)

Fun Applications

- Many of the web summarization tools aren't that interesting.
  - (You could probably build one just as good yourself after this lecture)
- Fun stuff happens when people take existing (boring) technology and apply it in a new (interesting) way.

AutoTLDR Bot

- www.reddit.com/user/autotldr
- Technology via smmry.com API
- Produces 3-sentence (3-paragraph?) summaries of articles posted to Reddit.
- Saves you a click

Facebook Is Draining Your iPhone Battery Because It's Tracking Your Location by **User_Name13** in technology

[–] **autotldr** 220 points 5 days ago

This is the best tl;dr I could make, original reduced by 79%. (I'm a bot)

> On Thursday, Circa co-founder Matt Gilligan wrote on Medium that the Facebook app on his iPhone was running in the background, even when he turned off background refresh for the app, which is supposed to shut off most app functions.
>
> Security researcher Jonathan Zdziarski analyzed the Facebook app's code and found that the app is sending devices' location information to Facebook in the background.
>
> Facebook's in-app location settings clearly state that if you have location history and access set to "Always," then Facebook "Will build a history of our precise location, even when you're not using the app."

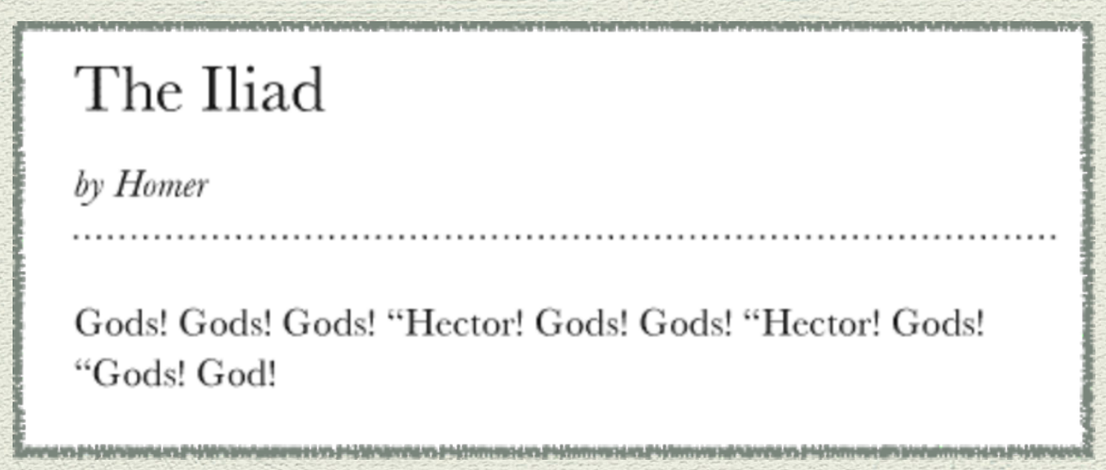
Extended Summary | FAQ | Theory | Feedback |
Top five keywords: **app**#1 **Facebook**#2 **location**#3 **background**#4 **set**#5

Post found in /r/technology and /r/Newsbeard.

permalink context full comments (918)

AutoSummarize

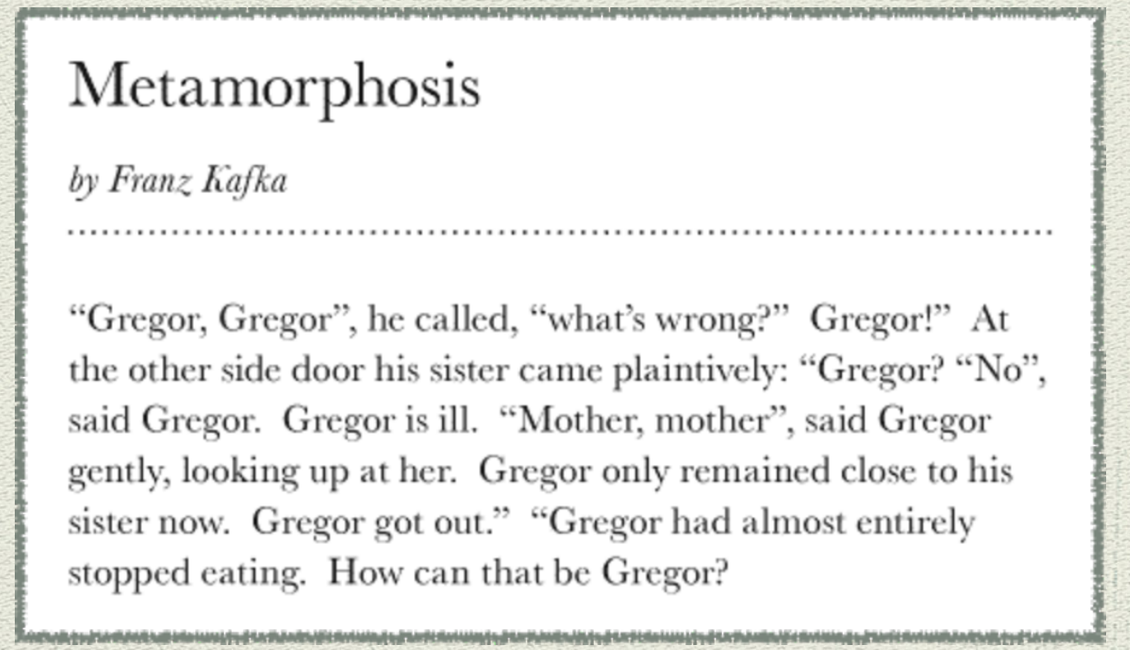
- By Jason Huff
- jason-huff.com/projects/autosummarize/
- Art project using MS Word 2008's *AutoSummarize* feature
- Compresses works into ten sentences
- Illustrates limitations of relevance approach



The Iliad

by Homer

Gods! Gods! Gods! "Hector! Gods! Gods! "Hector! Gods! "Gods! God!



Metamorphosis

by Franz Kafka

"Gregor, Gregor", he called, "what's wrong?" Gregor!" At the other side door his sister came plaintively: "Gregor? "No", said Gregor. Gregor is ill. "Mother, mother", said Gregor gently, looking up at her. Gregor only remained close to his sister now. Gregor got out." "Gregor had almost entirely stopped eating. How can that be Gregor?

Summarization Recap:

- Three-step method:
  1. Content Selection
  2. Information Ordering
  3. Sentence Realization
- Frequency method
- Graph method

- Extractive v Abstractive
- Generic v user-based
- Single-doc vs Multi-doc
- LLR (for topic words)
- MMR
- ROUGE
- SCUs
- Pyramid method