Interesting Phrase Mining

Alekh Jindal

Supervisors:

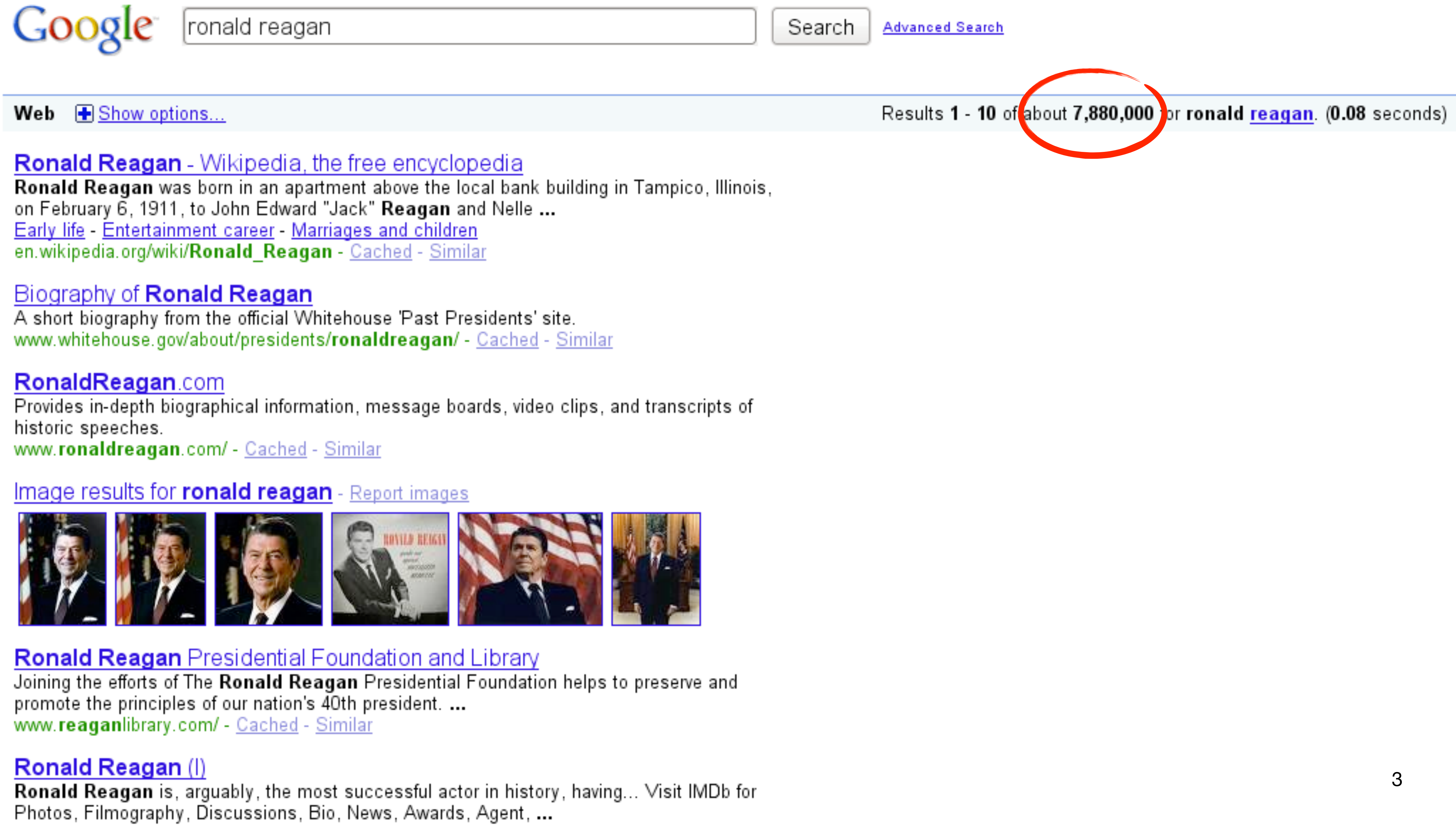
Prof. Jens Dittrich

Prof. Gerhard Weikum

What is a Phrase?

- A sequence of words intended to have a meaning - *Answers.com*

Why Phrase Mining?



Google ronald reagan Search Advanced Search

Web ✚ Show options... Results **1** - **10** of about **7,880,000** for **ronald reagan**. (**0.08** seconds)

Ronald Reagan - Wikipedia, the free encyclopedia
Ronald Reagan was born in an apartment above the local bank building in Tampico, Illinois, on February 6, 1911, to John Edward "Jack" **Reagan** and Nelle **...**
Early life - Entertainment career - Marriages and children
en.wikipedia.org/wiki/**Ronald_Reagan** - Cached - Similar

Biography of **Ronald Reagan**
A short biography from the official Whitehouse 'Past Presidents' site.
www.whitehouse.gov/about/presidents/**ronaldreagan**/ - Cached - Similar

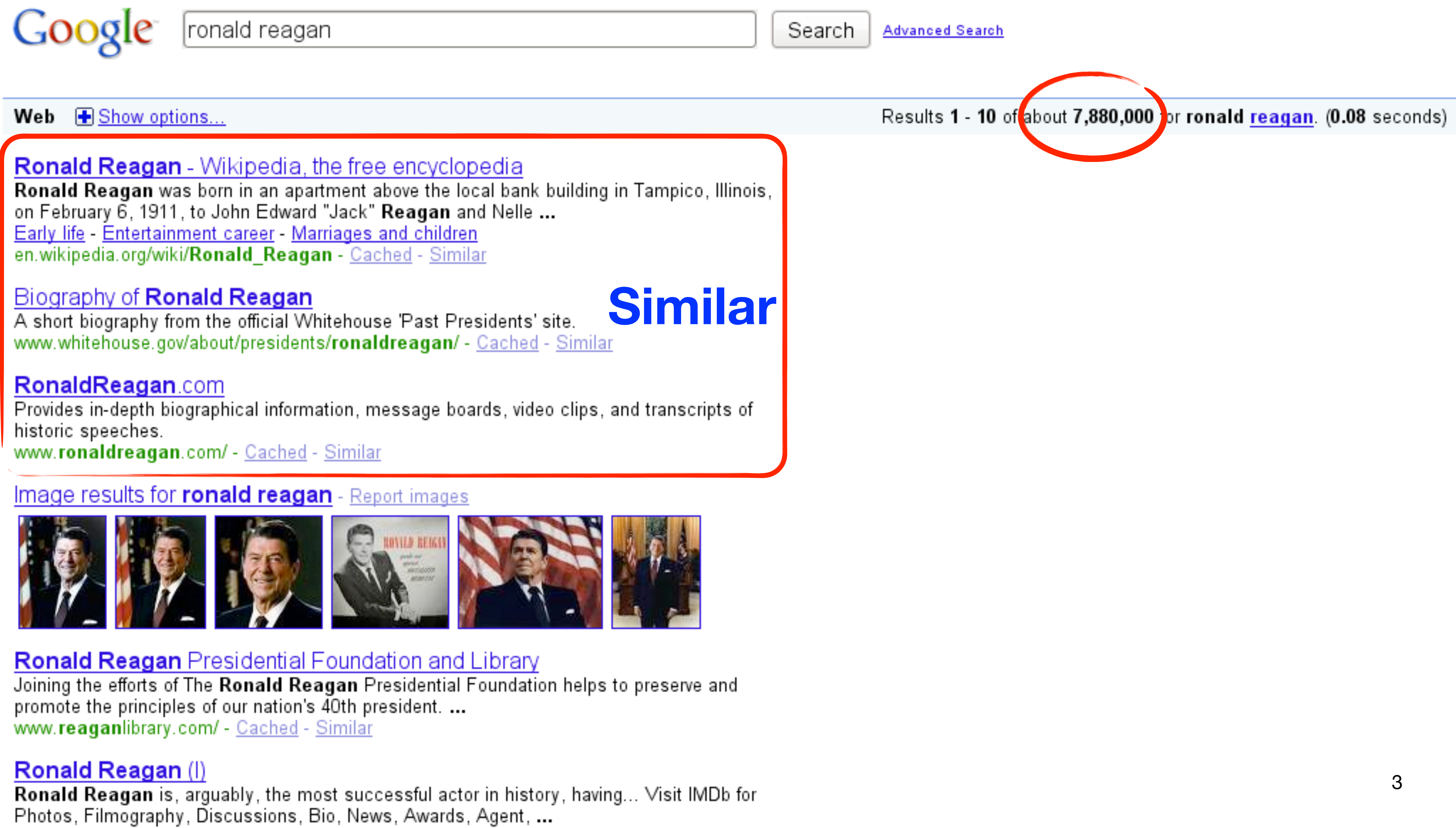
RonaldReagan.com
Provides in-depth biographical information, message boards, video clips, and transcripts of historic speeches.
www.**ronaldreagan**.com/ - Cached - Similar

Image results for **ronald reagan** - Report images

Ronald Reagan Presidential Foundation and Library
Joining the efforts of The **Ronald Reagan** Presidential Foundation helps to preserve and promote the principles of our nation's 40th president. **...**
www.**reagan**library.com/ - Cached - Similar

Ronald Reagan (I)
Ronald Reagan is, arguably, the most successful actor in history, having... Visit IMDb for Photos, Filmography, Discussions, Bio, News, Awards, Agent, **...**

Why Phrase Mining?



Google ronald reagan Search [Advanced Search]

Web Show options... Results **1** - **10** of about **7,880,000** for **ronald reagan**. (**0.08** seconds)

Ronald Reagan - Wikipedia, the free encyclopedia
Ronald Reagan was born in an apartment above the local bank building in Tampico, Illinois, on February 6, 1911, to John Edward "Jack" **Reagan** and Nelle **...**
Early life - Entertainment career - Marriages and children
en.wikipedia.org/wiki/**Ronald_Reagan** - Cached - Similar

Biography of **Ronald Reagan**
A short biography from the official Whitehouse 'Past Presidents' site.
www.whitehouse.gov/about/presidents/**ronaldreagan**/ - Cached - Similar

Similar

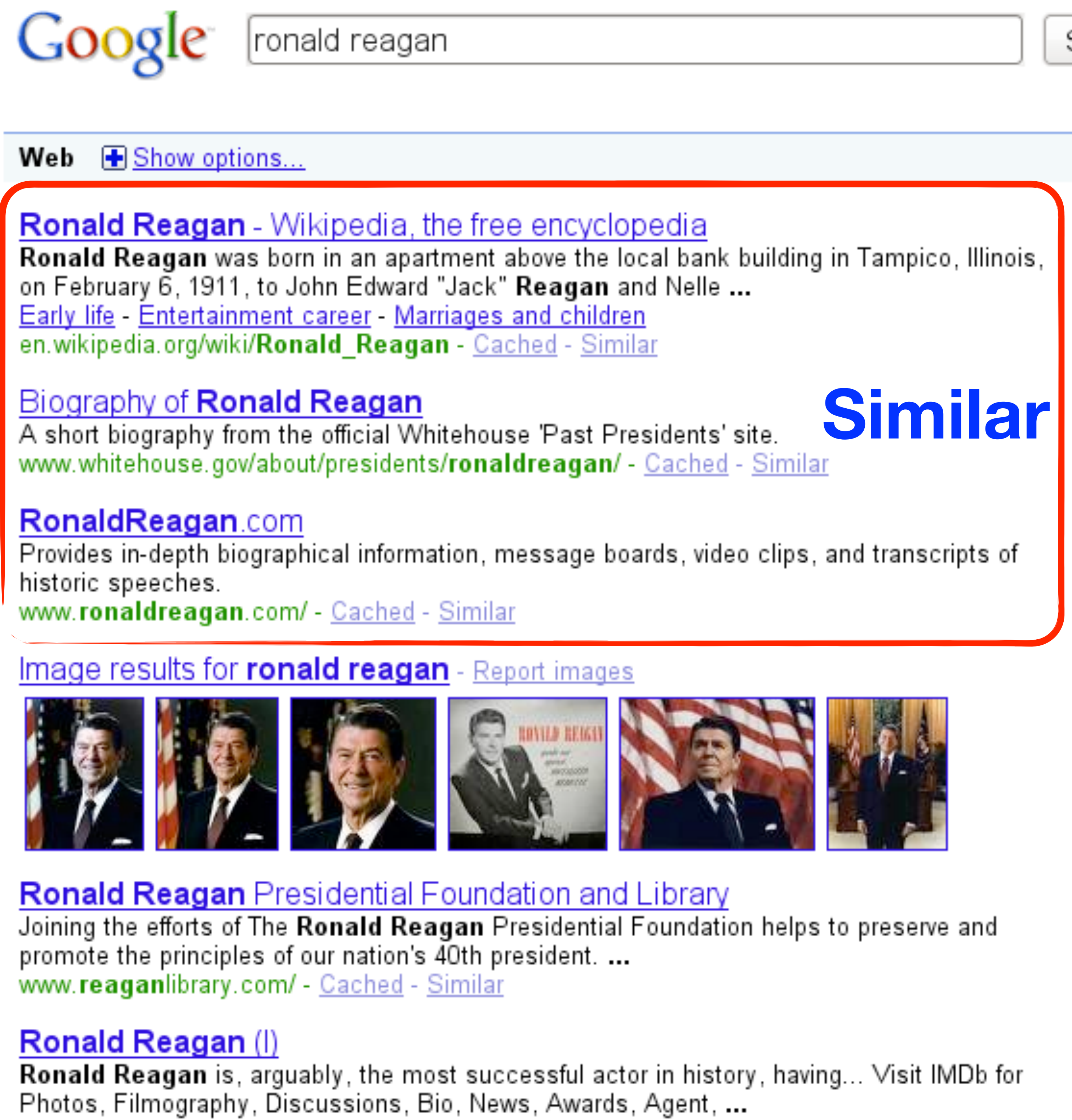
RonaldReagan.com
Provides in-depth biographical information, message boards, video clips, and transcripts of historic speeches.
www.**ronaldreagan**.com/ - Cached - Similar

Image results for **ronald reagan** - Report images

Ronald Reagan Presidential Foundation and Library
Joining the efforts of The **Ronald Reagan** Presidential Foundation helps to preserve and promote the principles of our nation's 40th president. **...**
www.**reagan**library.com/ - Cached - Similar

Ronald Reagan (I)
Ronald Reagan is, arguably, the most successful actor in history, having... Visit IMDb for Photos, Filmography, Discussions, Bio, News, Awards, Agent, **...**

Why Phrase Mining?



Google [ronald reagan] Search Advanced Search

Web Show options... Results **1** - **10** of about **7,880,000** for **ronald reagan**. (**0.08** seconds)

Ronald Reagan - Wikipedia, the free encyclopedia
Ronald Reagan was born in an apartment above the local bank building in Tampico, Illinois, on February 6, 1911, to John Edward "Jack" **Reagan** and Nelle **...**
Early life - Entertainment career - Marriages and children
en.wikipedia.org/wiki/**Ronald_Reagan** - Cached - Similar

Biography of **Ronald Reagan**
A short biography from the official Whitehouse 'Past Presidents' site.
www.whitehouse.gov/about/presidents/**ronaldreagan**/ - Cached - Similar

Similar

RonaldReagan.com
Provides in-depth biographical information, message boards, video clips, and transcripts of historic speeches.
www.**ronaldreagan**.com/ - Cached - Similar

Image results for **ronald reagan** - Report images

Ronald Reagan Presidential Foundation and Library
Joining the efforts of The **Ronald Reagan** Presidential Foundation helps to preserve and promote the principles of our nation's 40th president. **...**
www.**reagan**library.com/ - Cached - Similar

Ronald Reagan (I)
Ronald Reagan is, arguably, the most successful actor in history, having... Visit IMDb for Photos, Filmography, Discussions, Bio, News, Awards, Agent, **...**

Wish List

Political activities

Public statements

Personal life

Nick names

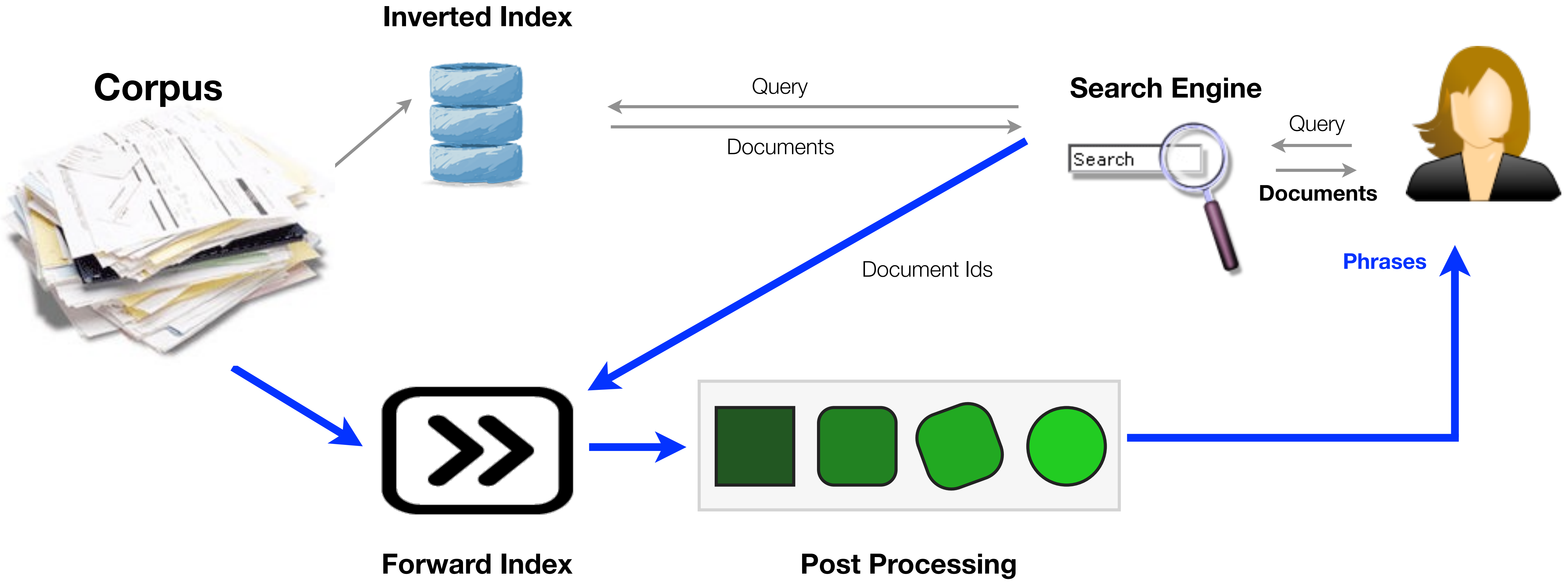
Anything else interesting about him?

What is an *Interesting* Phrase?

- Unique
- Complete
- Infrequent terms

Examples
- News: *obama wins nobel peace prize*
- Marketing slogans: *ithink therefore iMac*
- Quotations: *they that live in sin shall die in sin*

How does the system look like?



Inverted Index

Corpus

Query

Documents

Search Engine

Query

Documents

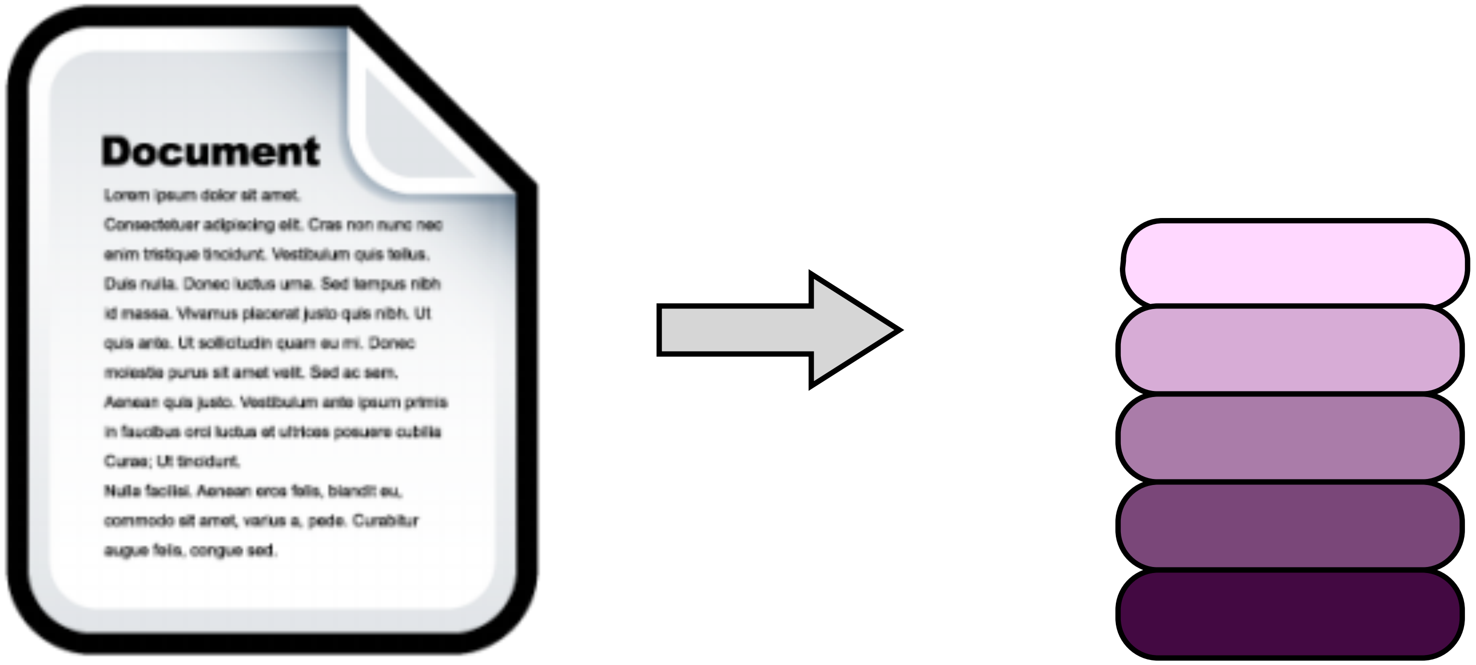
Phrases

Document Ids

Forward Index

Post Processing

How to get candidate set of Phrases?

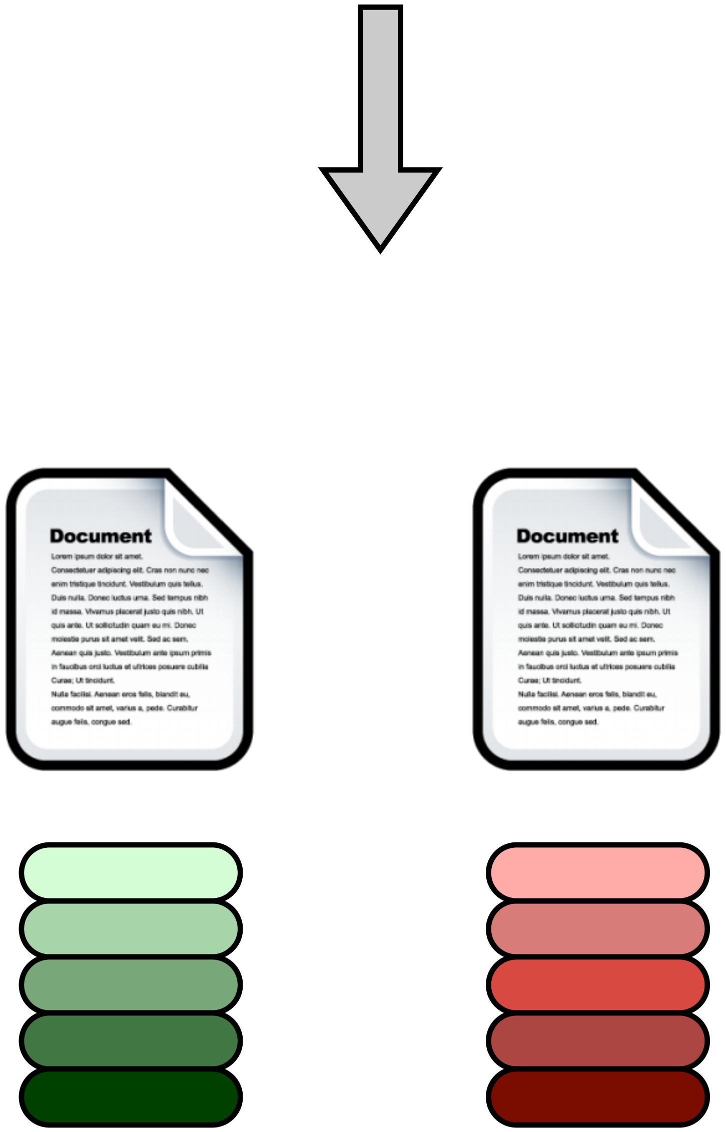


How to get candidate set of Phrases?

Document
Document
Document
Document

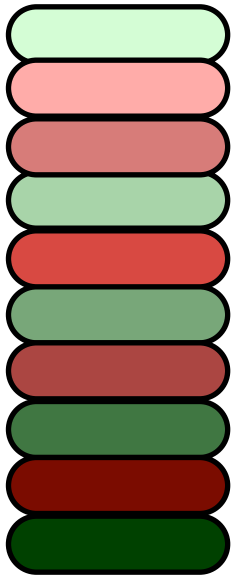
How to get candidate set of Phrases?

Query

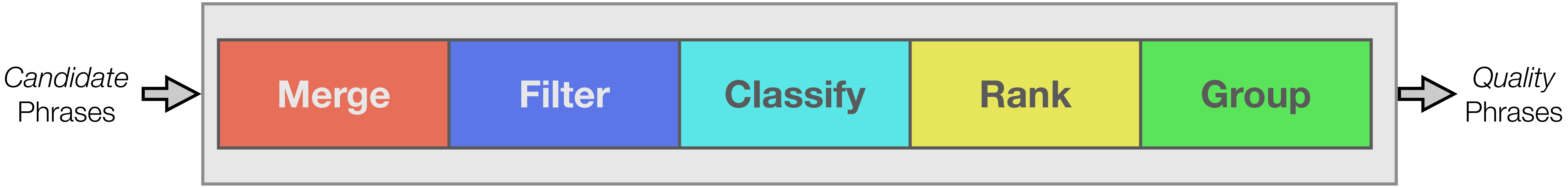


How to get candidate set of Phrases?

Candidate Phrases



Processing Pipeline



Candidate Phrases

Merge

Filter

Classify

Rank

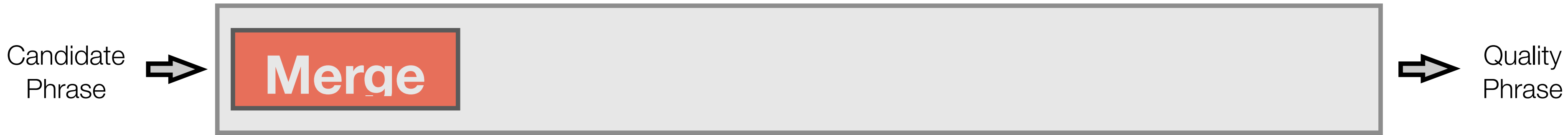
Group

Quality Phrases

Pipeline

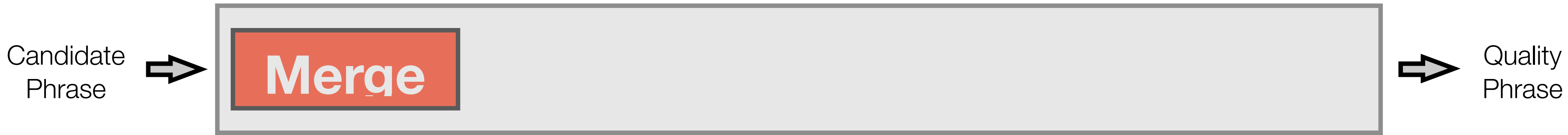
Phrase Merging

- Idea: Identify supplementing/complementing phrases
- Objective: Refine the candidate set of phrases to unique ones



Merge Strategies

- Exact merge
  - Sort phrases
  - Sliding window to merge matching phrases
- Approximate merge
  - Find edit distance between phrases
  - Merge phrases within distance threshold

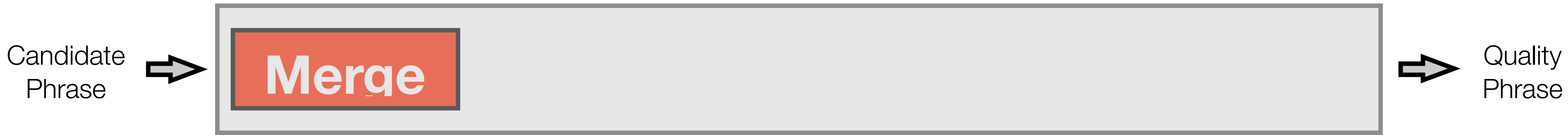


Exact Merge

- Prefix merge

united states captain tom

united states captain tom gorman



Exact Merge

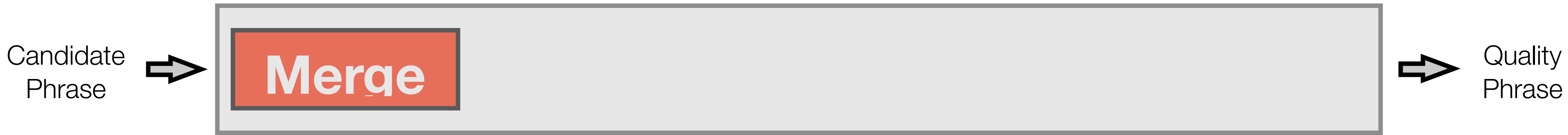
- Prefix merge

united states captain tom gorman

- Suffix merge

5 countries have ratified the kyoto

ratified the kyoto



Exact Merge

- Prefix merge

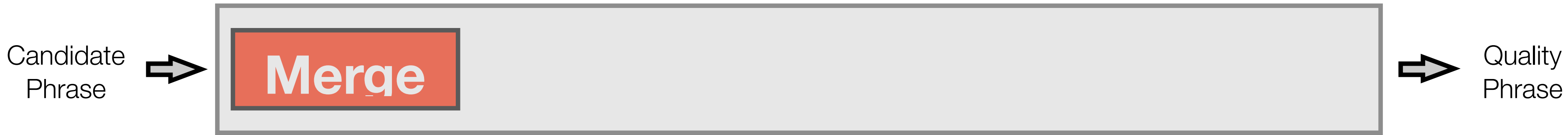
united states captain tom gorman

- Suffix merge

5 countries have ratified the kyoto

- Prefix-Suffix merge

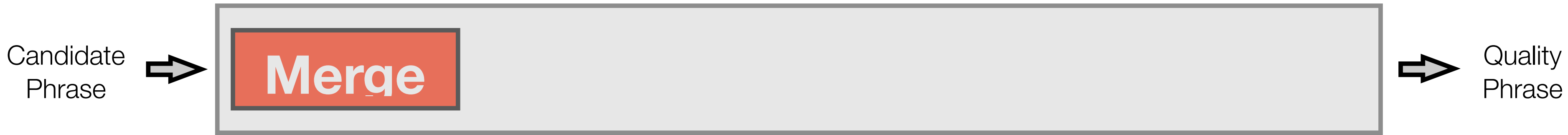
chief executive bollenbach announced   bollenbach announced annual results



Approximate Merge

- Stop-word merge

angela merkel chancellor $\xrightarrow{- \text{“the”}}$ angela merkel the chancellor



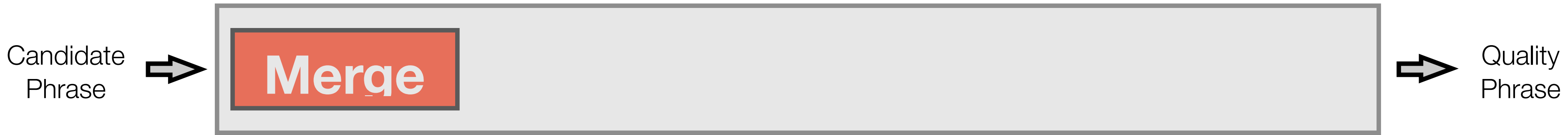
Approximate Merge

- Stop-word merge

angela merkel the chancellor

- Synonym merge

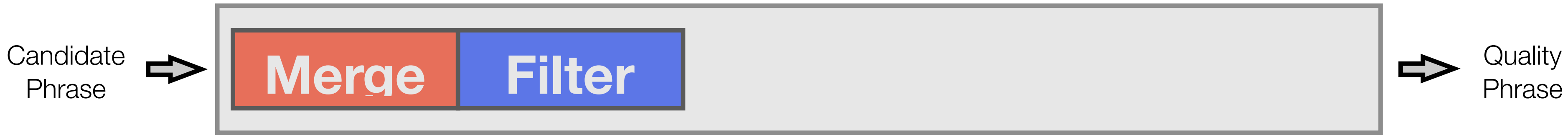
the jolly people — "jolly" / + "merry" → the merry people



Phrase Filtering

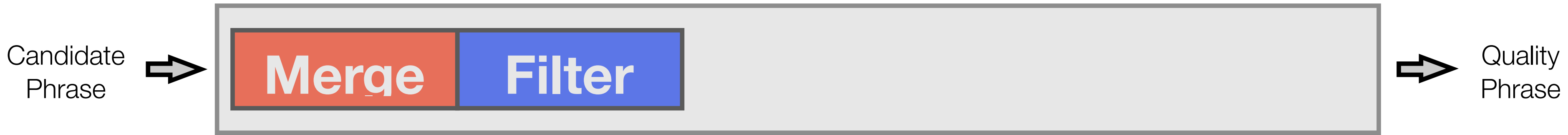
- Idea: Identify incomplete phrases

- Objective: Refine candidate set of phrases to meaningful ones



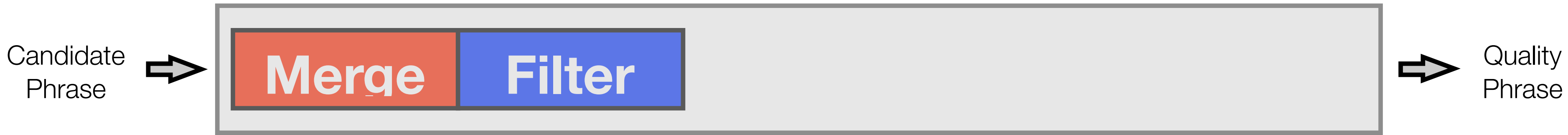
Filter Strategies

- Static rules based filter
- Fussy-Tree filter
- Classification filter



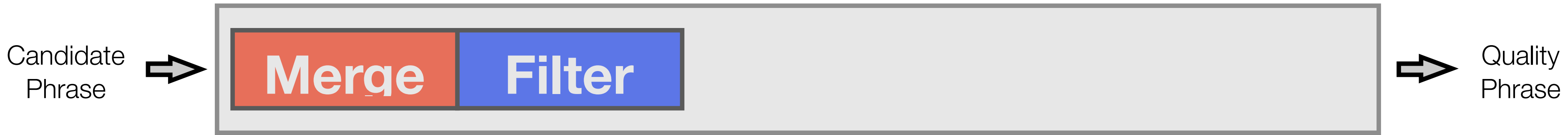
Static Rules based Filter

- Prefix/Suffix filter
  - phrases ending with “an”, “a”, “and”, “or”, “the”
- Parts-Of-Speech filter
  - certain sequences of parts of speech indicate incomplete phrases
  - e.g. actress jennifer lopez in care (noun) of (conjunction)



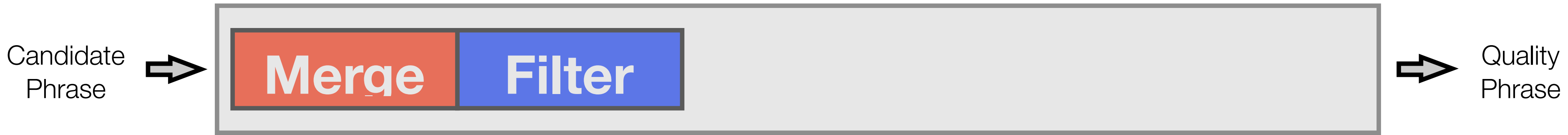
Fussy-Tree filter

- Phrase suffix tree (sparse) over the full corpus
- Only *significant* phrases inserted into the suffix tree
- A phrase is incomplete if Fussy-Tree returns a completion to it



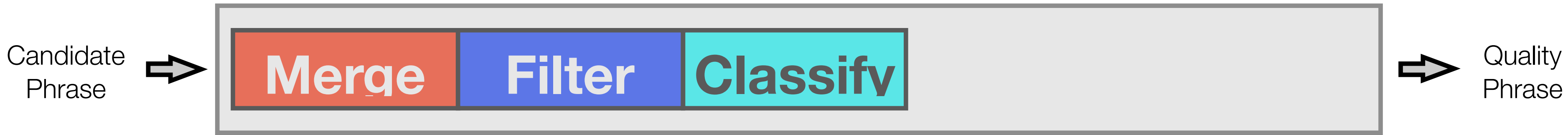
Classification Filter

- Aggressive filtering based on classifier estimates
- Filter out phrases classified as un-interesting with high confidence
- Applied post classification



Phrase Classification

- Feature extraction
- Training
- Classification


Candidate Phrase

Merge

Filter

Classify

Quality Phrase

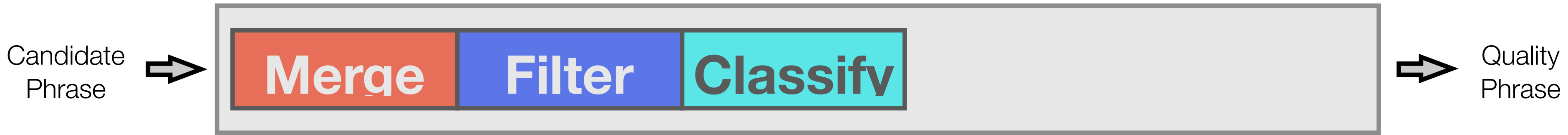
Feature Extraction

- Heuristics based
  e.g. number of terms, fresh terms, stop words ratio

- Frequency based
  e.g. inverse term frequencies

- Parts-of-Speech based
  e.g. nouns, adjectives, verbs

- Named entities based
  e.g. names of person, place, organization



Candidate Phrase

Merge

Filter

Classify

Quality Phrase

Training

- Labelled 1200 phrases over 12 queries
- Labels:

  3 - Very Interesting
  2 - Interesting
  1 - Not Good
  0 - Poor

Candidate Phrase

Merge

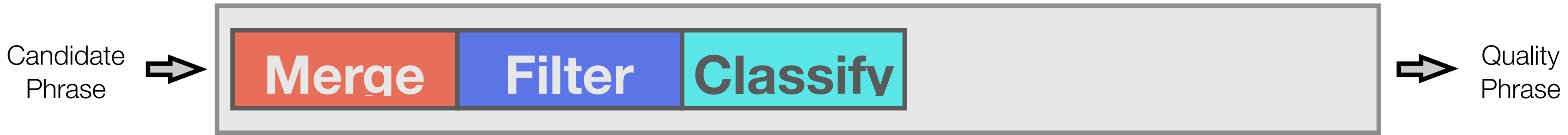
Filter

Classify

Quality Phrase

Classification

- Classification via regression (linear)
- Find line of best fit
- Use probability distributions of labels for ranking



Candidate Phrase

Merge

Filter

Classify

Quality Phrase

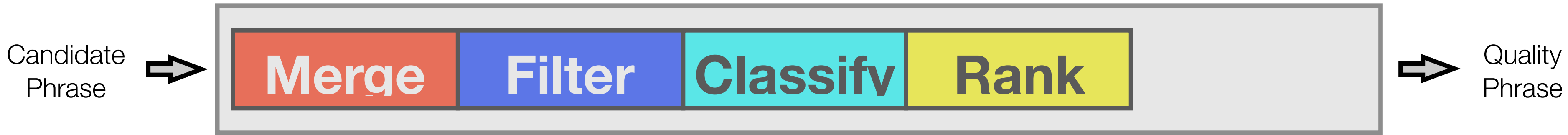
Phrase Ranking

- Use probability distribution of labels for ranking

- Model -

  Linear combination:

  Score = 2*P(label=3) + 1*P(label=2) - 1*P(label=1) - 2*P(label=0)


Candidate Phrase

Merge

Filter

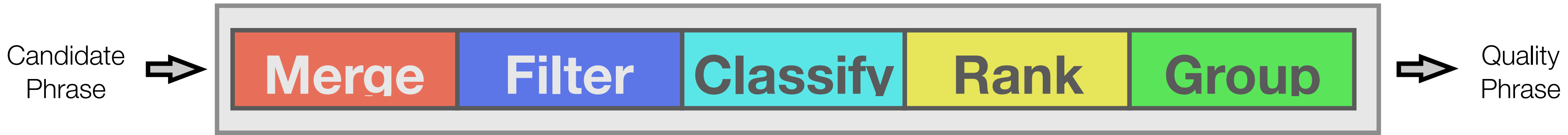
Classify

Rank

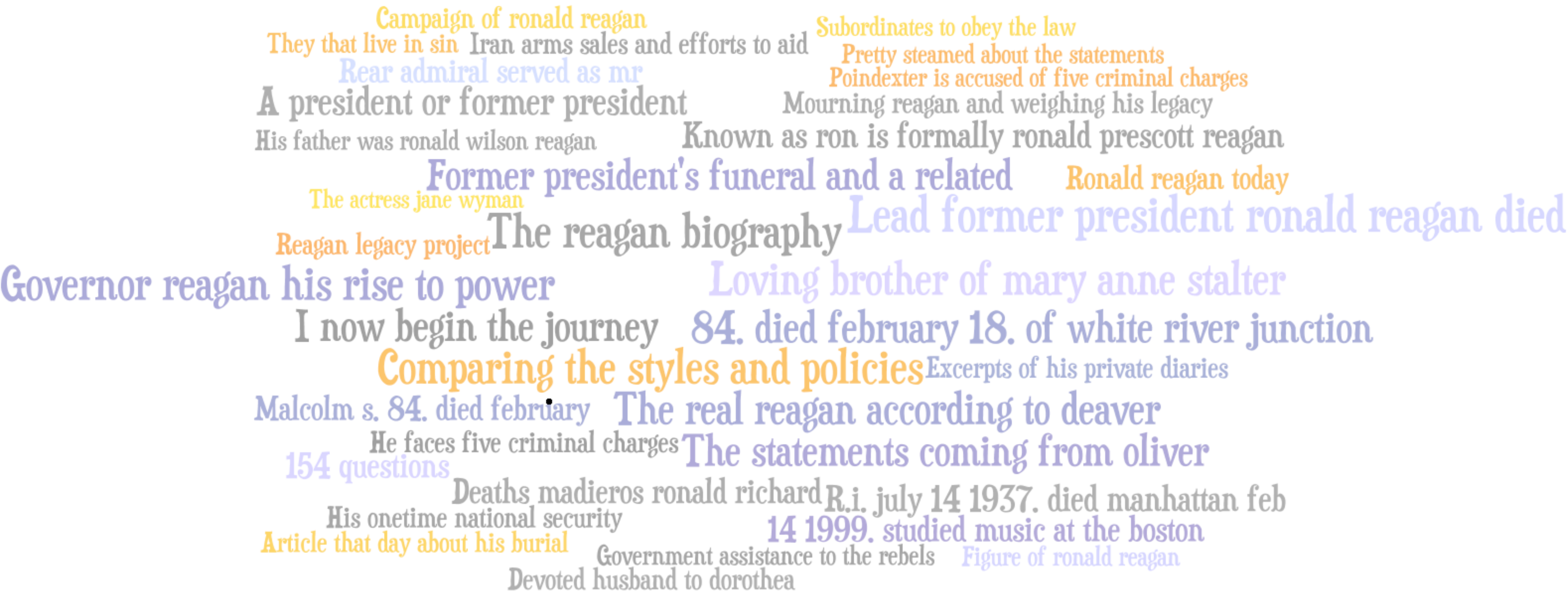
Quality Phrase

Phrase Grouping

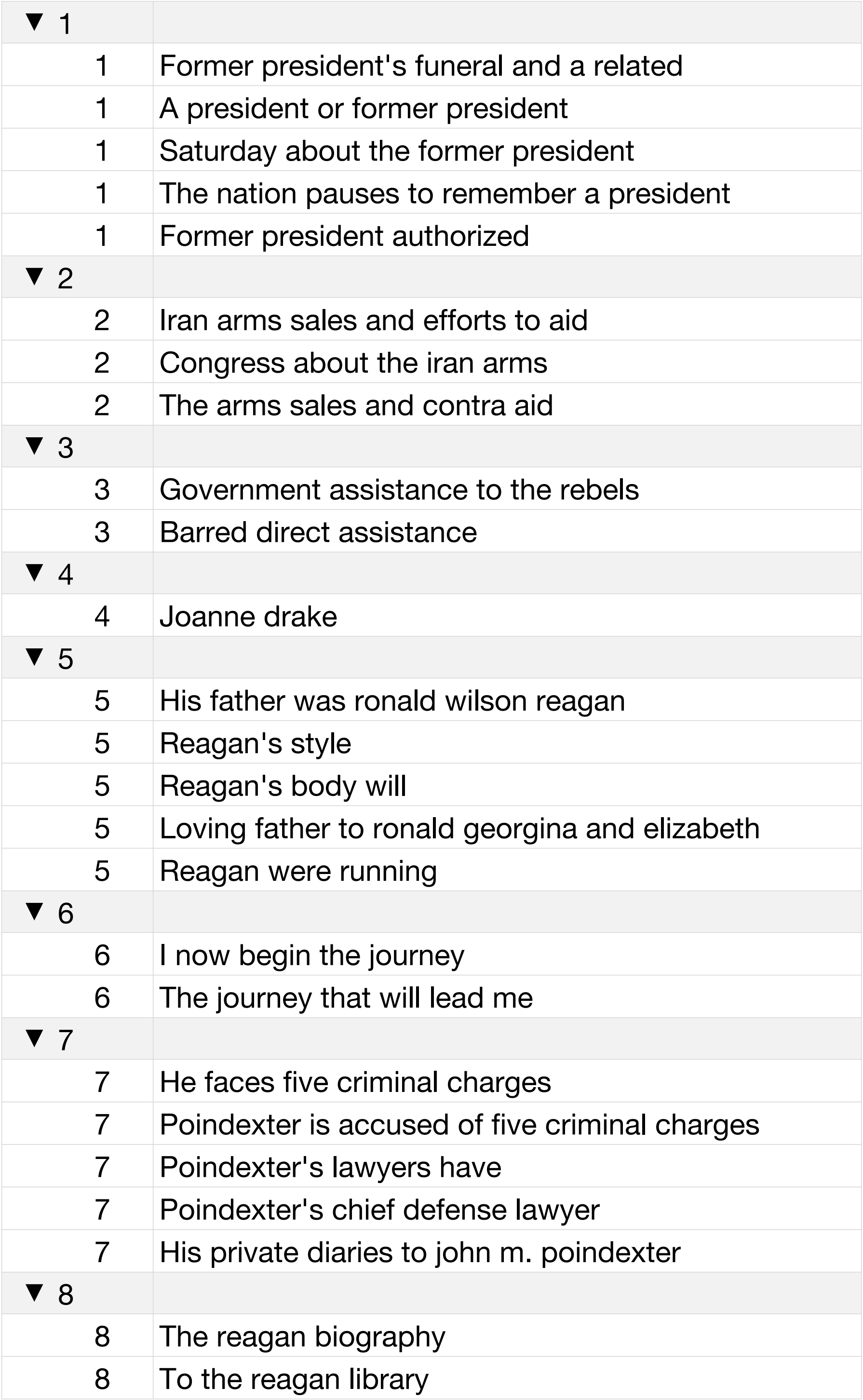
- Feature based
  - subset of features used for clustering
- Similarity based
  - nouns similarity
  - cosine similarity



Results


Campaign of ronald reagan
They that live in sin
Iran arms sales and efforts to aid
Subordinates to obey the law
Pretty steamed about the statements
Rear admiral served as mr
Poindexter is accused of five criminal charges
A president or former president
Mourning reagan and weighing his legacy
His father was ronald wilson reagan
Known as ron is formally ronald prescott reagan
Former president's funeral and a related
Ronald reagan today
The actress jane wyman
Lead former president ronald reagan died
Reagan legacy project
The reagan biography
Governor reagan his rise to power
Loving brother of mary anne stalter
I now begin the journey
84. died february 18. of white river junction
Comparing the styles and policies
Excerpts of his private diaries
Malcolm s. 84. died february
The real reagan according to deaver
He faces five criminal charges
154 questions
The statements coming from oliver
Deaths madieros ronald richard
R.i. july 14 1937. died manhattan feb
His onetime national security
14 1999. studied music at the boston
Article that day about his burial
Government assistance to the rebels
Figure of ronald reagan
Devoted husband to dorothea

Groups



▼ 1	
1	Former president's funeral and a related
1	A president or former president
1	Saturday about the former president
1	The nation pauses to remember a president
1	Former president authorized
▼ 2	
2	Iran arms sales and efforts to aid
2	Congress about the iran arms
2	The arms sales and contra aid
▼ 3	
3	Government assistance to the rebels
3	Barred direct assistance
▼ 4	
4	Joanne drake
▼ 5	
5	His father was ronald wilson reagan
5	Reagan's style
5	Reagan's body will
5	Loving father to ronald georgina and elizabeth
5	Reagan were running
▼ 6	
6	I now begin the journey
6	The journey that will lead me
▼ 7	
7	He faces five criminal charges
7	Poindexter is accused of five criminal charges
7	Poindexter's lawyers have
7	Poindexter's chief defense lawyer
7	His private diaries to john m. poindexter
▼ 8	
8	The reagan biography
8	To the reagan library

Evaluation Setup

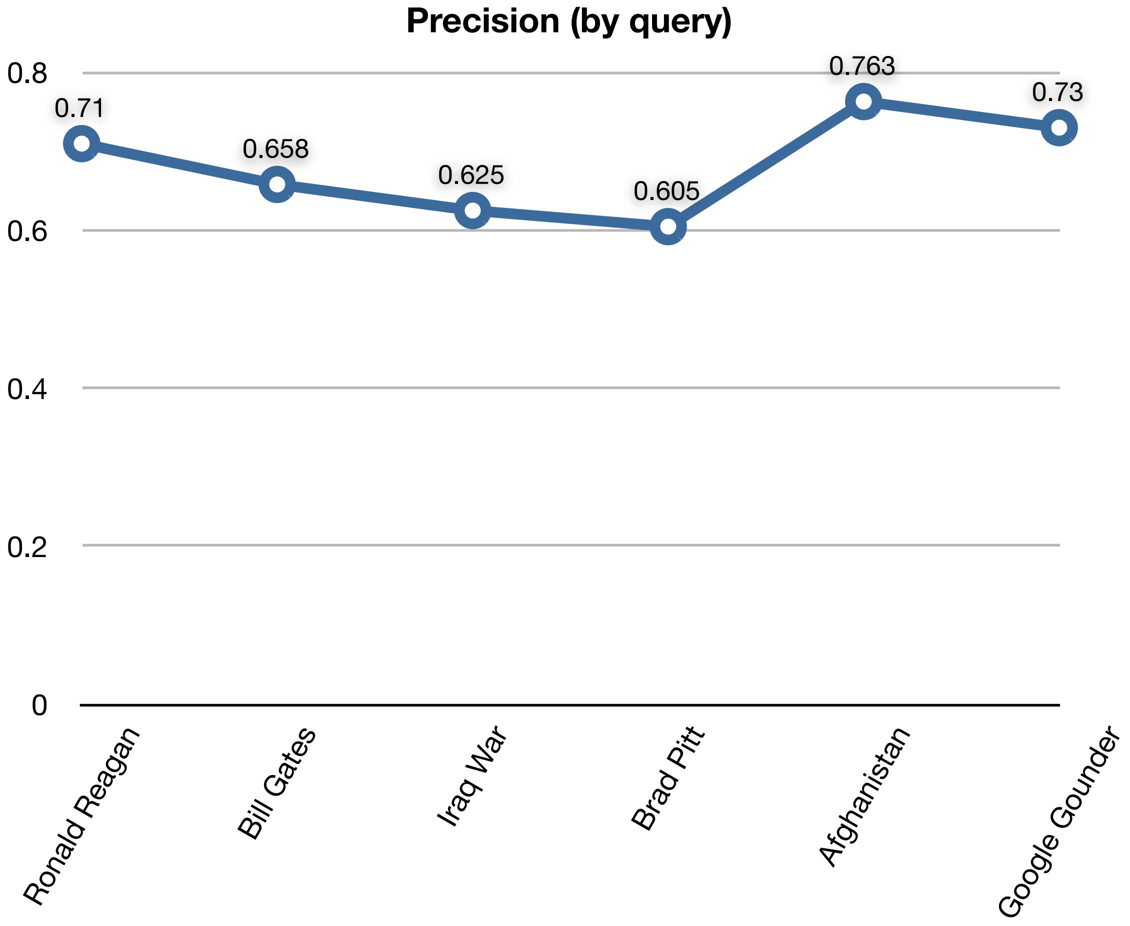
- 6 queries
  - Ronald Reagan, Bill Gates, Iraq War, Brad Pitt, Afghanistan, Google Founder

- Manually labelled candidate phrases

- Metrics
  - Precision
  - Recall
  - Time Latency
  - Filtering Effectiveness
  - Normalized Discounted Cumulative Gain (nDCG)

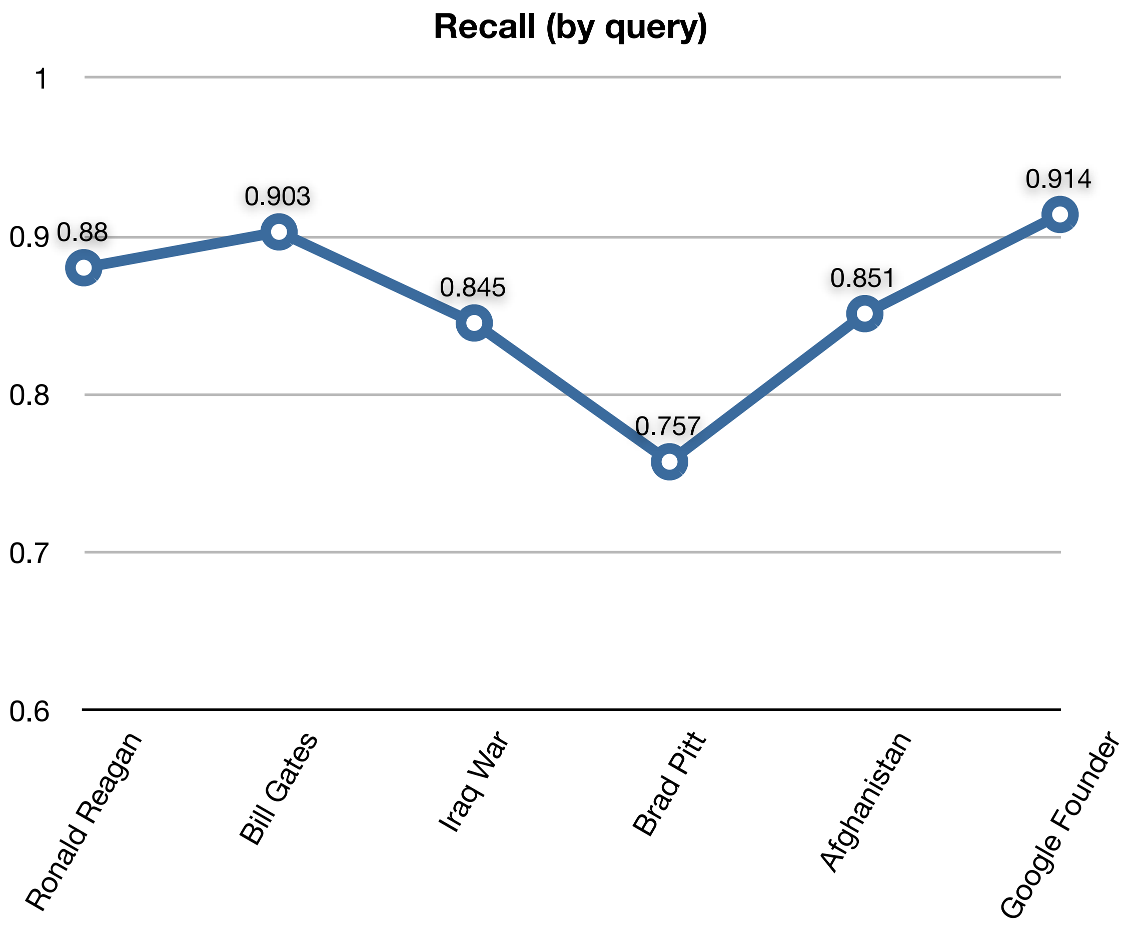
Precision

Q1	Q2	Q3	Q4	Q5	Q6
0.7096774194	0.6580645161	0.625	0.6046511628	0.7629310345	0.7298850575



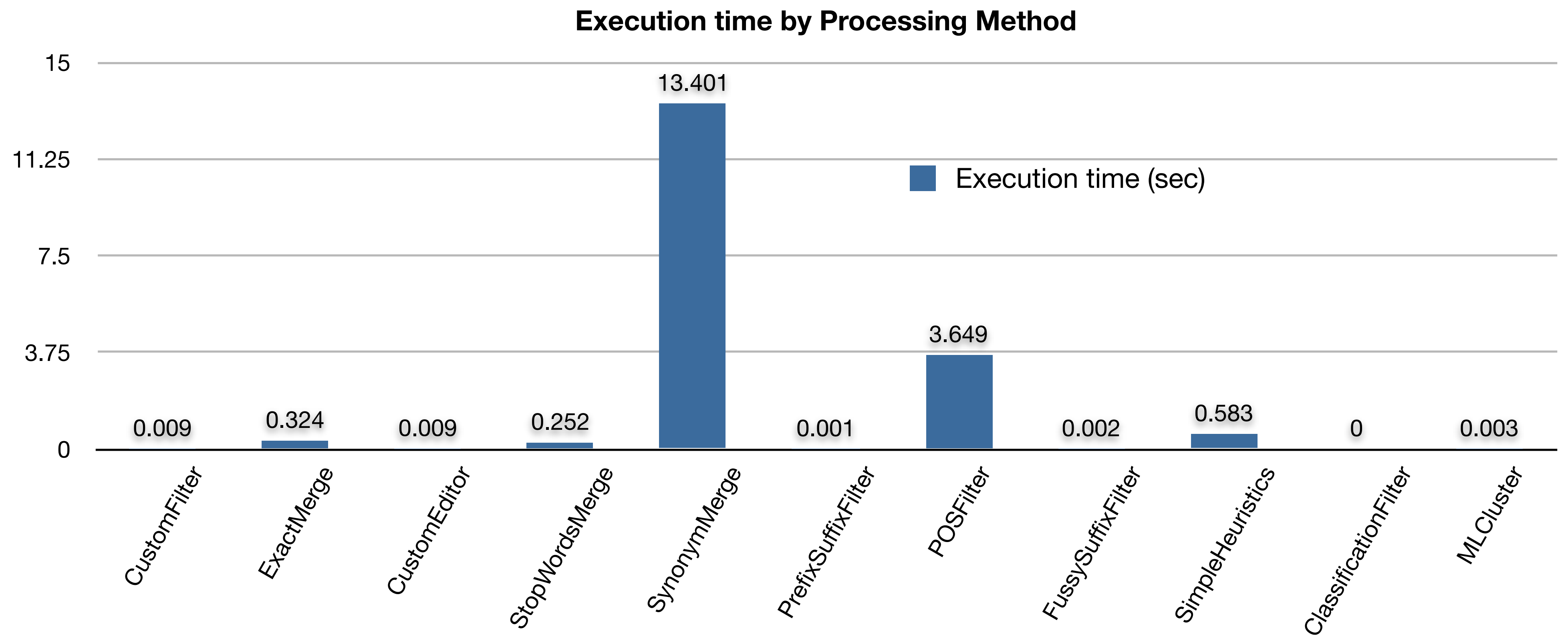
Recall

Q1	Q2	Q3	Q4	Q5	Q6
0.88	0.9026548673	0.8450704225	0.7572815534	0.8509615385	0.9136690647



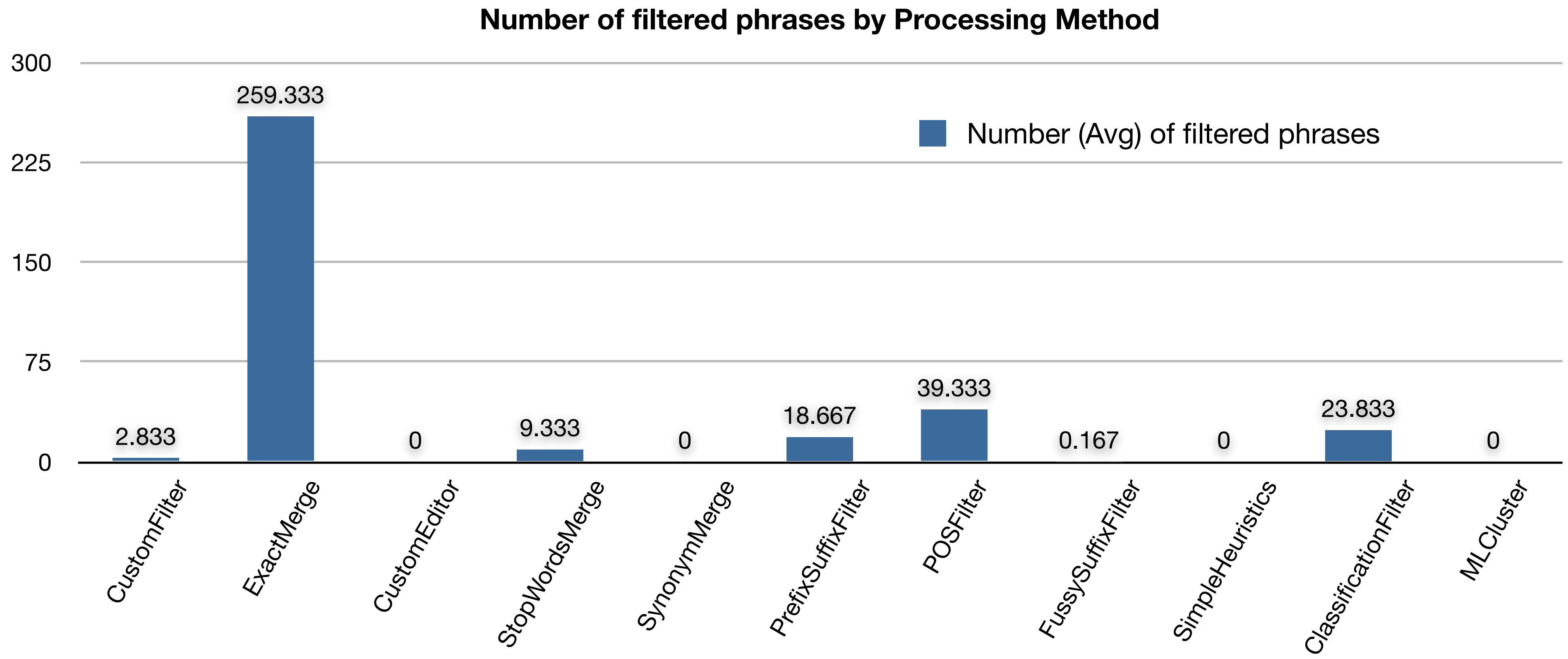
Time Latency

Method	Q1	Q1	Q3	Q4	Q5	Q6	Average
CustomFilter	0.027	0.004	0.003	0.003	0.003	0.015	0.0091667
ExactMerge	0.198	0.435	0.082	0.415	0.476	0.34	0.3243333
CustomEditor	0.028	0.007	0.013	0.002	0.004	0.002	0.0093333
StopWordMerge	0.414	0.294	0.068	0.173	0.294	0.27	0.2521667
SynonymMerge	11.752	6.621	11.056	22.414	17.118	11.442	13.4005
PrefixSuffixFilter	0.0010	0.0	0.0	0.0	0.001	0.0010	0.0005
POSFilter	5.462	3.269	2.23	3.622	5.108	2.2	3.6485
FussySuffixFilter	0.001	0.009	0.0010	0.0010	0.001	0.001	0.0023333
SimpleHeuristics	0.914	0.564	0.422	0.482	0.667	0.451	0.5833333
ClassificationFilter	0.0	0.0	0.0	0.0	0.0	0.0	0
MLCluster	0.0030	0.0040	0.0010	0.0030	0.0030	0.0010	0.0025



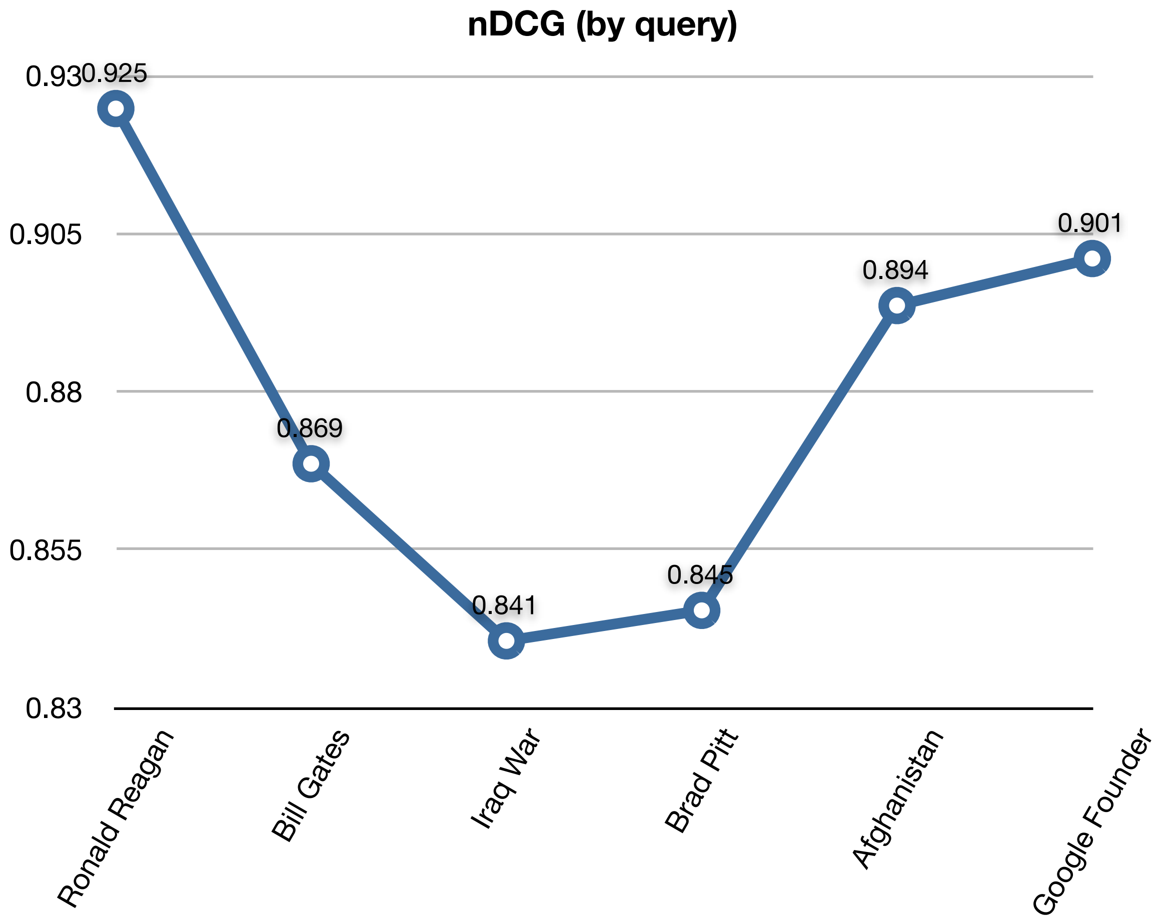
Filtering Effectiveness

Method	Q1	Q1	Q3	Q4	Q5	Q6	Average
CustomFilter	15	2	0	0	0	0	2.83333333
ExactMerge	326	241	350	287	121	231	259.333333
CustomEditor	0	0	0	0	0	0	0
StopWordMerge	8	19	0	0	23	6	9.33333333
SynonymMerge	0	0	0	0	0	0	0
PrefixSuffixFilter	13	18	13	20	29	19	18.6666667
POSFilter	22	36	30	49	59	40	39.3333333
FussySuffixFilter	0	0	0	0	1	0	0.16666667
SimpleHeuristics	0	0	0	0	0	0	0
ClassificationFilter	23	29	11	15	35	30	23.8333333
MLCluster	0	0	0	0	0	0	0



Normalized Discounted Cumulative Gain (nDCG)

Q1	Q2	Q3	Q4	Q5	Q6
0.92475894	0.8685600095	0.8405315749	0.8453590681	0.8935879424	0.9010311869



Conclusion

- Interesting phrases can offer first take insight of data
- Need to post-process phrases
- A combination of approaches work
- User study crucial
- It’s just a beginning!

Thanks!

- Questions?
- Remarks?
- Suggestions?