Interesting Phrase Mining

What is a Phrase?

Group of words functioning as a single unit in the syntax of a sentence.

Wikipedia

Motivation

Google paris hilton Search Advanced Search Preferences

Web Show options... Results **1** - **10** of about **65,600,000** for **paris hilton**. (**0.22** seconds)

News results for **paris hilton**

Gallery: **Paris Hilton**: Never a dull moment - 7 hours ago
Paris Hilton leaves federal court in Miami, Friday, July 10, 2009. **Hilton** is accused in a federal lawsuit of failing to adequately promote her 2006 movie **...**
Huntington Herald Dispatch - 1163 related articles »
Paris Hilton mourns former 'BFF' finalist - msnbc.com - 30 related articles »
Paris Hilton wears Georges Chakra for Lebanon appearance - iloubnan.info - 4 related articles »

Paris Hilton - Wikipedia, the free encyclopedia
Paris Whitney **Hilton** (born February 17, 1981) is an American socialite, media personality, model, singer, author and actress. **...**
en.wikipedia.org/wiki/**Paris_Hilton** - Cached - Similar

Paris Hilton | The Official Website
ParisHilton.com **Paris Hilton**, Nicky **Hilton** Fashion, Pictures, Apparel, Jewellery, Film, and Fun.
www.**parishilton**.com/ - Cached - Similar

Paris Hilton
Socialite **Paris Hilton** was born on February 17, 1981 in New York City... Visit IMDb for Photos, Filmography, Discussions, Bio, News, Awards, Agent, **...**
www.imdb.com/name/nm0385296/ - Cached - Similar

Image results for **paris hilton** - Report images

Video results for **paris hilton**

PARIS IN JAIL: The Music Video
2 min 47 sec
www.youtube.com

NEW ! **Paris Hilton** Sex Video!
7 min
video.google.com

Paris Hilton Zone | **Paris Hilton** Pictures, Pics, Photos
4000+ new **Paris Hilton** pictures, **Paris** wallpaper, sex tape, lyrics, audio, video, daily **Paris** pics & news.
Video - Pictures - Wallpaper - TV
www.**parishilton**zone.com/ - Cached - Similar

MySpace.com - **Paris Hilton** - 28 - Female - Beverly Hills **...**

Motivation

Google paris hilton Search Advanced Search Preferences

Web Show options... Results 1 - 10 of about 65,600,000 for paris hilton. (0.22 seconds)

News results for **paris hilton**

Gallery: **Paris Hilton**: Never a dull moment - 7 hours ago
Paris Hilton leaves federal court in Miami, Friday, July 10, 2009. **Hilton** is accused in a federal lawsuit of failing to adequately promote her 2006 movie **...**
Huntington Herald Dispatch - 1163 related articles »
Paris Hilton mourns former 'BFF' finalist - msnbc.com - 30 related articles »
Paris Hilton wears Georges Chakra for Lebanon appearance - iloubnan.info - 4 related articles »

Paris Hilton - Wikipedia, the free encyclopedia
Paris Whitney **Hilton** (born February 17, 1981) is an American socialite, media personality, model, singer, author and actress. **...**
en.wikipedia.org/wiki/**Paris_Hilton** - Cached - Similar

Paris Hilton | The Official Website
ParisHilton.com **Paris Hilton**, Nicky **Hilton** Fashion, Pictures, Apparel, Jewellery, Film, and Fun.
www.**parishilton**.com/ - Cached - Similar

Paris Hilton
Socialite **Paris Hilton** was born on February 17, 1981 in New York City... Visit IMDb for Photos, Filmography, Discussions, Bio, News, Awards, Agent, **...**
www.imdb.com/name/nm0385296/ - Cached - Similar

Similar

Image results for **paris hilton** - Report images

Video results for **paris hilton**

PARIS IN JAIL: The Music Video
2 min 47 sec
www.youtube.com

NEW ! **Paris Hilton** Sex Video!
7 min
video.google.com

Paris Hilton Zone | **Paris Hilton** Pictures, Pics, Photos
4000+ new **Paris Hilton** pictures, **Paris** wallpaper, sex tape, lyrics, audio, video, daily **Paris** pics & news.
Video - Pictures - Wallpaper - TV
www.**parishilton**zone.com/ - Cached - Similar

MySpace.com - **Paris Hilton** - 28 - Female - Beverly Hills **...**

Motivation

Google paris hilton Search Advanced Search Preferences

Web Show options... Results 1 - 10 of about 65,600,000 for paris hilton. (0.22 seconds)

News results for **paris hilton**

Gallery: **Paris Hilton**: Never a dull moment - 7 hours ago
Paris Hilton leaves federal court in Miami, Friday, July 10, 2009. **Hilton** is accused in a federal lawsuit of failing to adequately promote her 2006 movie **...**
Huntington Herald Dispatch - 1163 related articles »
Paris Hilton mourns former 'BFF' finalist - msnbc.com - 30 related articles »
Paris Hilton wears Georges Chakra for Lebanon appearance - iloubnan.info - 4 related articles »

Paris Hilton - Wikipedia, the free encyclopedia
Paris Whitney **Hilton** (born February 17, 1981) is an American socialite, media personality, model, singer, author and actress. **...**
en.wikipedia.org/wiki/**Paris_Hilton** - Cached - Similar

Paris Hilton | The Official Website
ParisHilton.com **Paris Hilton**, Nicky **Hilton** Fashion, Pictures, Apparel, Jewellery, Film, and Fun.
www.**parishilton**.com/ - Cached - Similar

Paris Hilton
Socialite **Paris Hilton** was born on February 17, 1981 in New York City... Visit IMDb for Photos, Filmography, Discussions, Bio, News, Awards, Agent, **...**
www.imdb.com/name/nm0385296/ - Cached - Similar

Image results for **paris hilton** - Report images

Video results for **paris hilton**

PARIS IN JAIL: The Music Video
2 min 47 sec
www.youtube.com

NEW ! **Paris Hilton** Sex Video!
7 min
video.google.com

Paris Hilton Zone | **Paris Hilton** Pictures, Pics, Photos
4000+ new **Paris Hilton** pictures, **Paris** wallpaper, sex tape, lyrics, audio, video, daily **Paris** pics & news.
Video - Pictures - Wallpaper - TV
www.**parishilton**zone.com/ - Cached - Similar

MySpace.com - **Paris Hilton** - 28 - Female - Beverly Hills **...**

Similar

Wish List

- Family
- Personal Info
- Albums
- Hotel Business
- Reckless driving
- Recent news

What are phrases of interest?

- Names (people, organizations, products)

  “sarah palin”, “amnesty international”, “iphone”

- News headlines

  “non aligned movement leaders meet”

- Marketing slogans

  “ithink therefore imac”

- Activities

  “reckless driving”

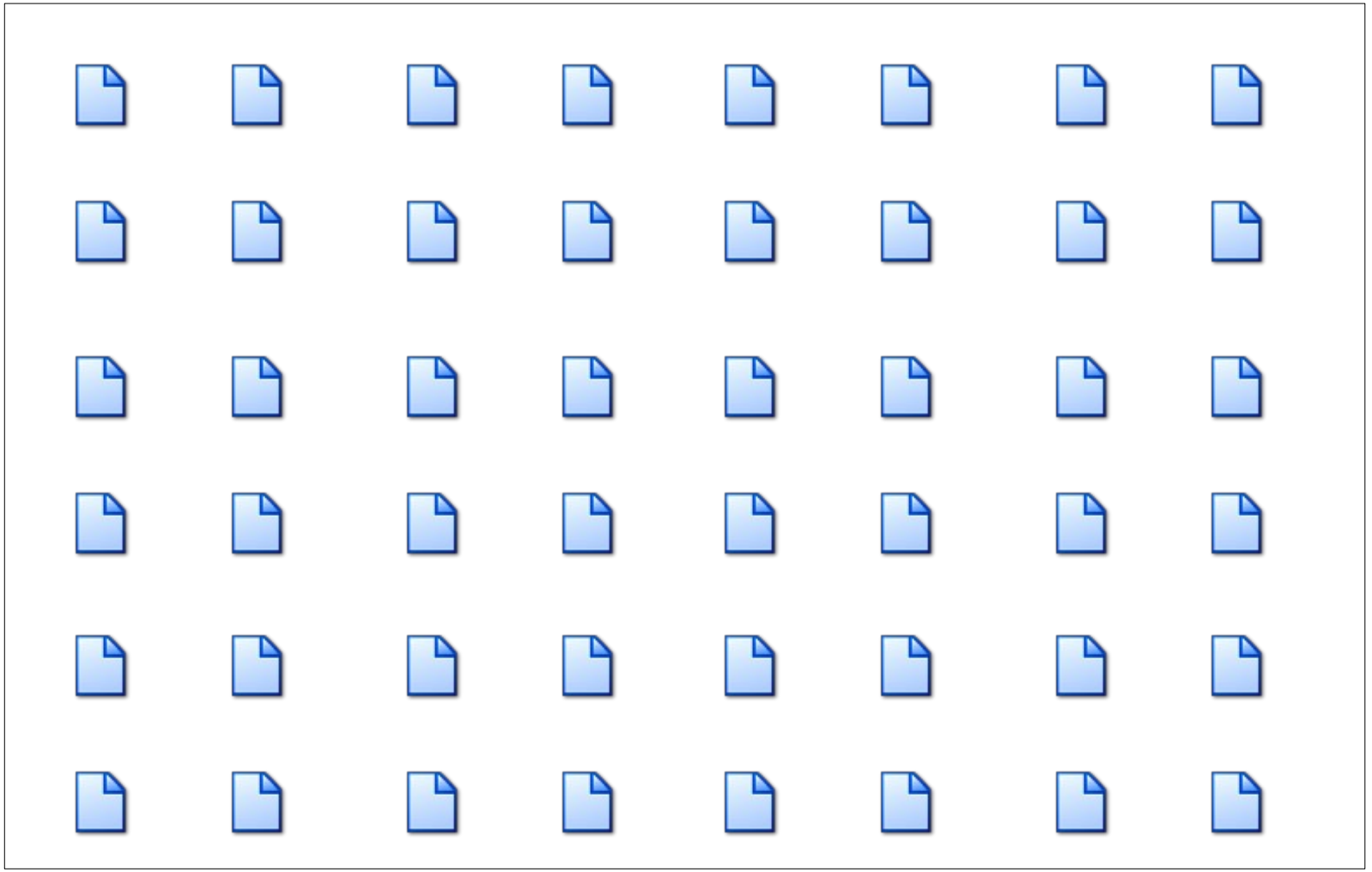
- Quotations

  “to be or not to be”

- .. and more

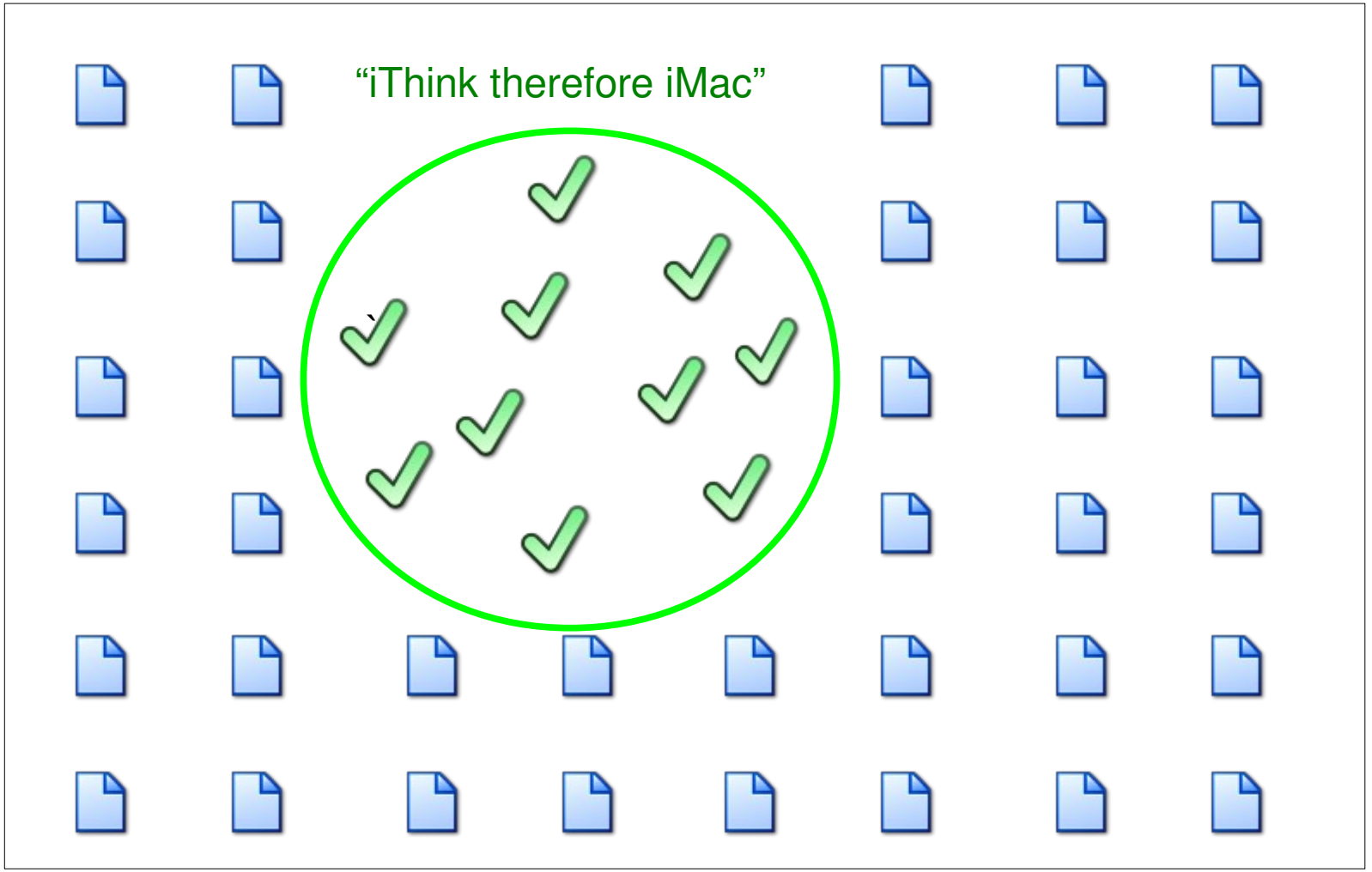
Definition of Interestingness

Corpus

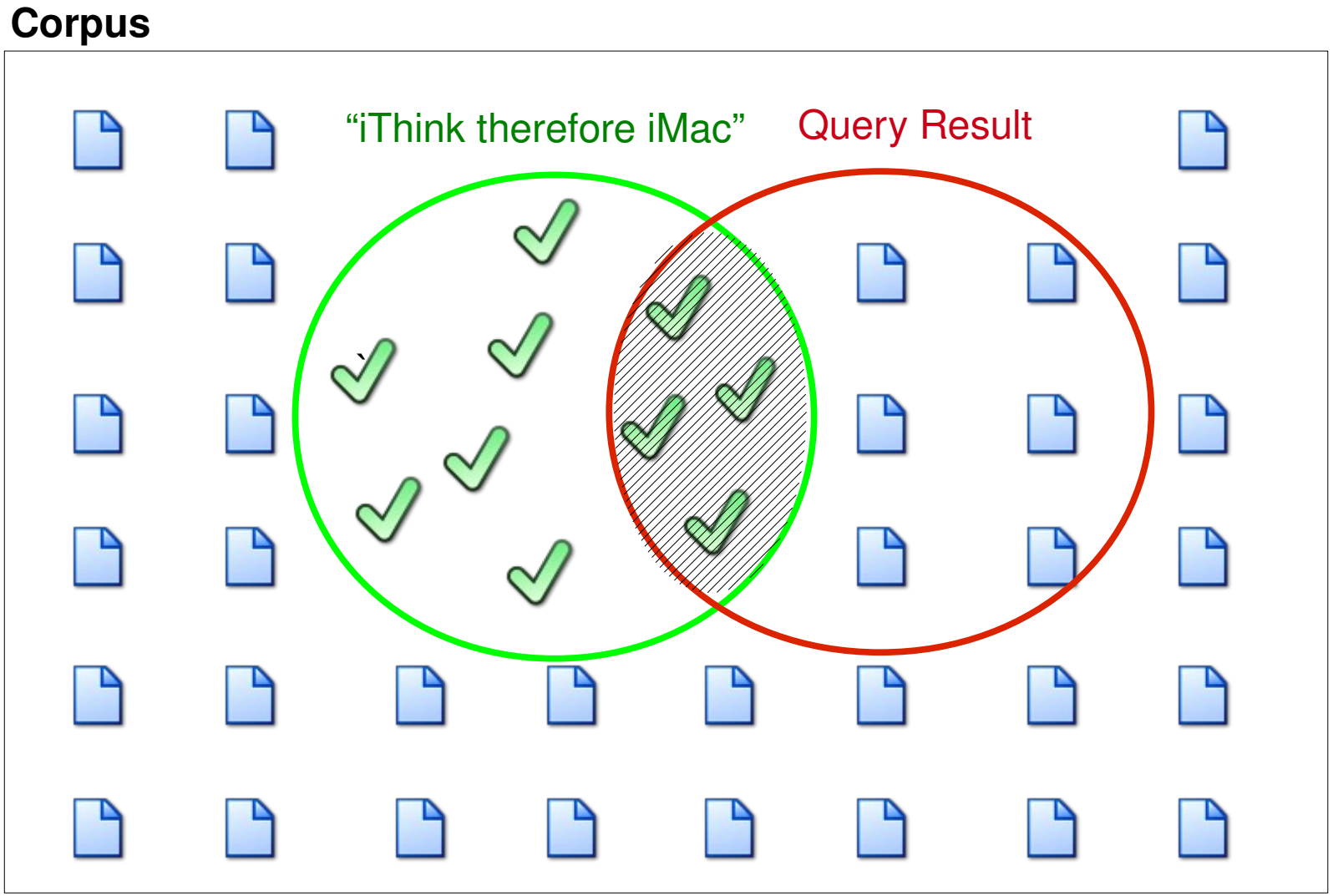


Definition of Interestingness

Corpus



Definition of Interestingness



Definition of Interestingness

Formally defined as:

$$I_D(p, D') = \frac{freq(p, D')}{freq(p, D)}$$

Where

- p is the phrase
- D is the document corpus
- D' is the result set (sub-collection of D)

Indexing and Query processing

- Prior work to efficiently index and query
- Phrases indexed using forward index
- The system retrieves a candidate set of phrases

16th July, 2009

Alekh Jindal


Which ones are relevant?

Post Processing

- Merge
- Group
- Classify
- Rank

Merge

- Merge phrases having exact matches
- Remove obvious duplicates
- Reduce candidate set of phrases

Merge Methods

- Prefix merge
- Suffix merge
- Prefix-suffix merge

Prefix Merge

Merge phrases which are prefix of other phrases

united states captain tom

united states captain tom gorman

united states captain tom gorman

Suffix Merge

Merge phrases which are suffix of other phrases

ratified the kyoto

5 countries have ratified the kyoto

5 countries have ratified the kyoto

Prefix-Suffix Merge

Merge phrases with common suffix and prefix respectively

hilton's chief executive bollenbach announced

bollenbach announced annual results

hilton's chief executive bollenbach announced annual results

Group

- Approximate merge “similar” phrases
- Remove implied duplicates
- Reduce candidate set of phrases

Grouping Methods

- Edit distance
- Synonyms
- Latent semantic analysis
- Simrank

Edit Distance

Minimum number of operations to transform one string into another

Merge	Group		

Edit Distance

S I T T I N G K I T T E N

Merge	Group		

Edit Distance

S I T T I N G

K I T T E N

Merge | Group

Edit Distance

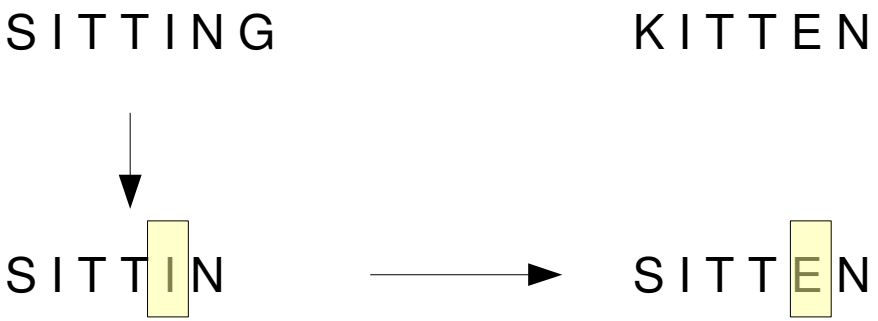
S I T T I N G

K I T T E N

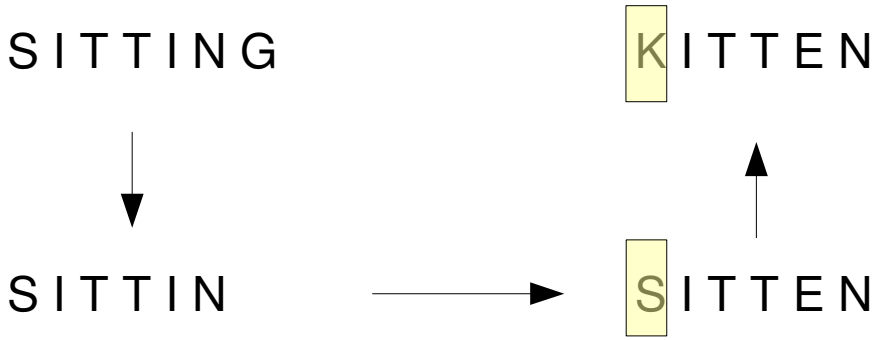
S I T T I N

Merge	Group		

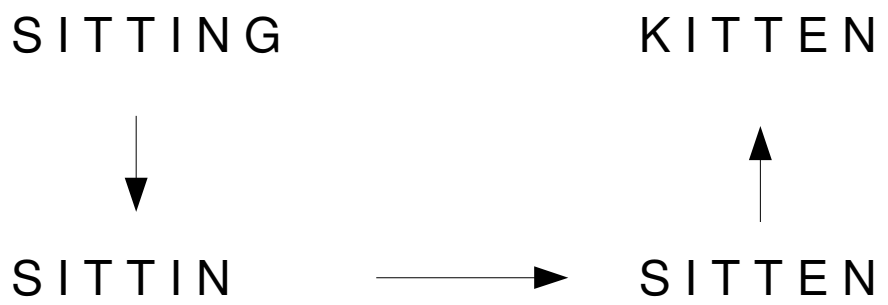
Edit Distance



Edit Distance



Edit Distance



Distance = 3

Edit Distance

Matrix Representation

		K	I	T	T	E	N
S							
I							
T							
T							
I							
N							
G							

d[i, j] = min operations for

Row[1…i] → Col[1...j]

Edit Distance

Matrix Representation

		K	I	T	T	E	N
	0	1	2	3	4	5	6
S							
I							
T							
T							
I							
N							
G							

Merge | Group

Edit Distance

Matrix Representation

		K	I	T	T	E	N
	0	1	2	3	4	5	6
S	1						
I	2						
T	3						
T	4						
I	5						
N	6						
G	7						

Merge | Group

Edit Distance

Matrix Representation

		K	I	T	T	E	N
	0	1	2	3	4	5	6
S	1	1					
I	2						
T	3						
T	4						
I	5						
N	6						
G	7						

Merge | Group

Edit Distance

Matrix Representation

		K	I	T	T	E	N
	0	1	2	3	4	5	6
S	1	1					
I	2	2					
T	3						
T	4						
I	5						
N	6						
G	7						

S I → K

Edit Distance

Matrix Representation

		K	I	T	T	E	N
	0	1	2	3	4	5	6
S	1	1	2	3	4	5	6
I	2	2	1	2	3	4	5
T	3	3	2	1	2	3	4
T	4	4	3	2	1	2	3
I	5	5	4	3	2	2	3
N	6	6	5	4	3	3	2
G	7	7	6	5	4	4	3

Edit Distance

Matrix Representation

		K	I	T	T	E	N
	0	1	2	3	4	5	6
S	1	1	2	3	4	5	6
I	2	2	1	2	3	4	5
T	3	3	2	1	2	3	4
T	4	4	3	2	1	2	3
I	5	5	4	3	2	2	3
N	6	6	5	4	3	3	2
G	7	7	6	5	4	4	3

Edit Distance

Edit Distance

Matrix Representation

		K	I	T	T	E	N
	0	1	2	3	4	5	6
S	1	1	2	3	4	5	6
I	2	2	1	2	3	4	5
T	3	3	2	1	2	3	4
T	4	4	3	2	1	2	3
I	5	5	4	3	2	2	3
N	6	6	5	4	3	3	2
G	7	7	6	5	4	4	3

d[i, j] = min (
d[i-1, j] + 1,
d[i, j-1] + 1,
d[i-1, j-1] + **cost**,
)

cost = 0, if Row[i] = Col[j]
= 1, otherwise

Edit Distance

Edit Distance

Applying to Phrases

angela merkel chancellor

angela merkel the chancellor

Merge	Group		

Edit Distance

Applying to Phrases

angela merkel chancellor

angela merkel the chancellor

Edit Distance

- Idea: Assign fractional weights to stop words
- Cost of inserting/deleting stop-word = 0.1

Edit Distance

- Idea: Assign fractional weights to stop words
- Cost of inserting/deleting stop-word = 0.1

```
d[i, j] = min (
            d[i-1, j] + 0.1,
            d[i, j-1] + 0.1,
            d[i-1, j-1] + cost
            )
```

Edit Distance

- Idea: Assign fractional weights to stop words
- Cost of inserting/deleting stop-word = 0.1

		angela	merkel	the	chancellor
	0	1	2		
angela	1	0	1		
merkel	2	1	0		
chancellor	3	2	1		

Edit Distance

- Idea: Assign fractional weights to stop words
- Cost of inserting/deleting stop-word = 0.1

		angela	merkel	the	chancellor
	0	1	2	2.1	3.1
angela	1	0	1	1.1	2.1
merkel	2	1	0	0.1	1.1
chancellor	3	2	1	1	0.1

Edit Distance

- Idea: Assign fractional weights to stop words
- Cost of inserting/deleting stop-word = 0.1

		angela	merkel	the	chancellor
	0	1	2	2.1	3.1
angela	1	0	1	1.1	2.1
merkel	2	1	0	0.1	1.1
chancellor	3	2	1	1	0.1

Synonyms

Idea: Consider synonym equality rather than exact equality while calculating edit distance

Merge	Group		

Synonyms

Idea: Consider synonym equality rather than exact equality while calculating edit distance

		four	got	killed
	0	1	2	3
four	1	0	1	2
people	2	1	1	2
died	3	2	2	1

Grouping – other methods

- Latent semantic analysis
  - Idea: Use term co-occurrence to extract similarity
- Simrank
  - Idea: Two phrases are similar if they are contained in similar documents
  - Graph theoretic idea

Classify

- Mark phrases as “relevant” or “irrelevant”
- Analogy: spam filter
- Reduce candidate set of phrases

Classify

- Feature generation
- Feature selection
- Classification

Feature Generation

Features: characteristics of phrases

Merge	Group	*Classify*	

Feature Generation

Heuristics: common sense rules

Merge	Group	*Classify*	

Feature Generation

Heuristics

- Length of phrase
- Number of articles
- Number of stop words
- Position of stop words
- Number of query terms

Feature Generation

Named entities

Feature Generation

Named entities

- Names of persons e.g. Angela Merkel
- Names of places e.g. Saarbruecken
- Names of organizations e.g. Saarland University

Feature Generation

Named entities

- Names of persons e.g. Angela Merkel
- Names of places e.g. Saarbruecken
- Names of organizations e.g. Saarland University

Note: Named entities extraction requires full sentences; needs to be done at indexing level.

Feature Generation

Parts of speech

Feature Generation

Parts of speech

- Nouns
- Verbs
- Adjectives
- Adverbs
- etc

Feature Generation

Parts of speech

- Nouns
- Verbs
- Adjectives
- Adverbs
- etc

Note: Parts of speech tagging requires full sentences; needs to be done at indexing level.

Feature Selection

- Idea: Choose the relevant features
- Use X^2 statistics
- Use training set

X^2 Statistics

Idea:

How far are feature and class from independence

X^2 Statistics

Example:

- Class labels: Relevant, Irrelevant
- Feature: Number of verbs

X^2 Statistics

Example:

- Class labels: Relevant, Irrelevant
- Feature: Number of verbs

	Feature: #Verbs = 0	Feature: #Verbs > 0
Class = relevant	2	90
Class = irrelevant	3	5

X^2 Statistics

Example:

- Class labels: Relevant, Irrelevant
- Feature: Number of verbs

N = 2+3+90+5 = 100

	Feature: #Verbs = 0	Feature: #Verbs > 0
Class = relevant	2	90
Class = irrelevant	3	5

X^2 Statistics

Example:

- Class labels: Relevant, Irrelevant
- Feature: Number of verbs

	Feature: #Verbs = 0	Feature: #Verbs > 0
Class = relevant	2	90
Class = irrelevant	3	5

N = 2+3+90+5 = 100

Independence Assumption:

P(#Verbs=0, Class=relevant)
= P(#Verbs=0)*P(Class=relevant)

X^2 Statistics

Example:

- Class labels: Relevant, Irrelevant
- Feature: Number of verbs

	Feature: #Verbs = 0	Feature: #Verbs > 0
Class = relevant	2	90
Class = irrelevant	3	5

N = 2+3+90+5 = 100

Independence Assumption:

P(#Verbs=0, Class=relevant)
= P(#Verbs=0)*P(Class=relevant)

P(#Verbs=0) = (2+3)/100 = 0.05

X^2 Statistics

Example:

- Class labels: Relevant, Irrelevant
- Feature: Number of verbs

	Feature: #Verbs = 0	Feature: #Verbs > 0
Class = relevant	2	90
Class = irrelevant	3	5

N = 2+3+90+5 = 100

Independence Assumption:

P(#Verbs=0, Class=relevant)
= P(#Verbs=0)*P(Class=relevant)

P(#Verbs=0) = (2+3)/100 = 0.05

P(Class=relevant) = (2+90)/100
= 0.92

X^2 Statistics

Example:

- Class labels: Relevant, Irrelevant
- Feature: Number of verbs

	Feature: #Verbs = 0	Feature: #Verbs > 0
Class = relevant	2	90
Class = irrelevant	3	5

N = 2+3+90+5 = 100

Independence Assumption:

P(#Verbs=0, Class=relevant)
= P(#Verbs=0)*P(Class=relevant)

P(#Verbs=0) = (2+3)/100 = 0.05

P(Class=relevant) = (2+90)/100
= 0.92

Estimated Occurrences:
= N.P(#Verb=0).P(Class=relevant)

X^2 Statistics

Example:

- Class labels: Relevant, Irrelevant
- Feature: Number of verbs

	Feature: #Verbs = 0	Feature: #Verbs > 0
Class = relevant	2 (4.6)	90
Class = irrelevant	3	5

N = 2+3+90+5 = 100

Independence Assumption:

P(#Verbs=0, Class=relevant)
= P(#Verbs=0)*P(Class=relevant)

P(#Verbs=0) = (2+3)/100 = 0.05

P(Class=relevant) = (2+90)/100
= 0.92

Estimated Occurrences:
= N.P(#Verb=0).P(Class=relevant)
= 4.6

X^2 Statistics

Example:
- Class labels: Relevant, Irrelevant
- Feature: Number of verbs

	Feature: #Verbs = 0	Feature: #Verbs > 0
Class = relevant	2 (4.6)	90 (87.4)
Class = irrelevant	3 (0.4)	5 (7.6)

N = 2+3+90+5 = 100

Independence Assumption:

P(#Verbs=0, Class=relevant)
= P(#Verbs=0)*P(Class=relevant)

P(#Verbs=0) = (2+3)/100 = 0.05

P(Class=relevant) = (2+90)/100
= 0.92

Estimated Occurrences:
= N.P(#Verb=0).P(Class=relevant)
= 4.6

X^2 Statistics

$$X^2(c, f) = \sum_{f_i} \frac{(f_o - f_e)^2}{f_e}$$

X^2 Statistics

$$X^2(c, f) = \sum_{f_i} \frac{(f_o - f_e)^2}{f_e}$$

$$X^2(c, f) = \frac{(2-4.6)^2}{4.6} + \frac{(3-0.4)^2}{0.4} + \frac{(90-87.4)^2}{87.4} + \frac{(5-7.6)^2}{7.6}$$

$$= 19.35$$

X^2 Statistics

$$X^2(c, f) = \sum_{f_i} \frac{(f_o - f_e)^2}{f_e}$$

$$X^2(c, f) = \frac{(2-4.6)^2}{4.6} + \frac{(3-0.4)^2}{0.4} + \frac{(90-87.4)^2}{87.4} + \frac{(5-7.6)^2}{7.6}$$

$$= 19.35$$

The **lower** the X^2 score, the **better** the feature characterizes the phrase

Classification via Regression

- Idea: Use training set to find decision boundary
- Linear regression: find line of best fit (least squares)

Classification via Regression

- Idea: Use training set to find decision boundary
- Linear regression: find line of best fit (least squares)

$$f(x; w) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

Classification via Regression

- Idea: Use training set to find decision boundary
- Linear regression: find line of best fit (least squares)

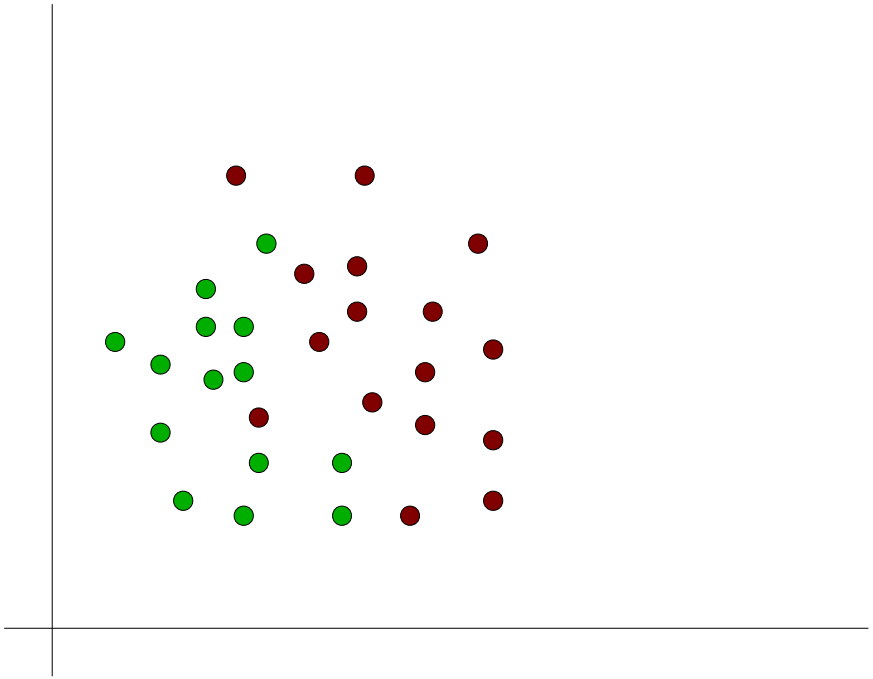
$$f(x;w) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

If f(x;w)>0.5, label = 1 (relevant)

Else, label = 0 (irrelevant)

Post Processing - Classify

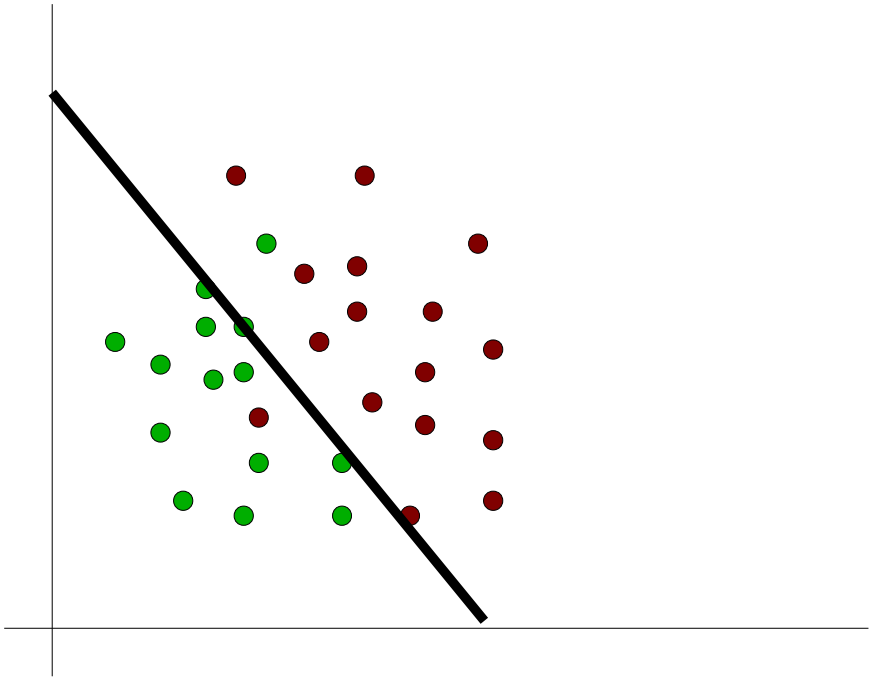
- Linear Regression



Merge	Group	*Classify*	

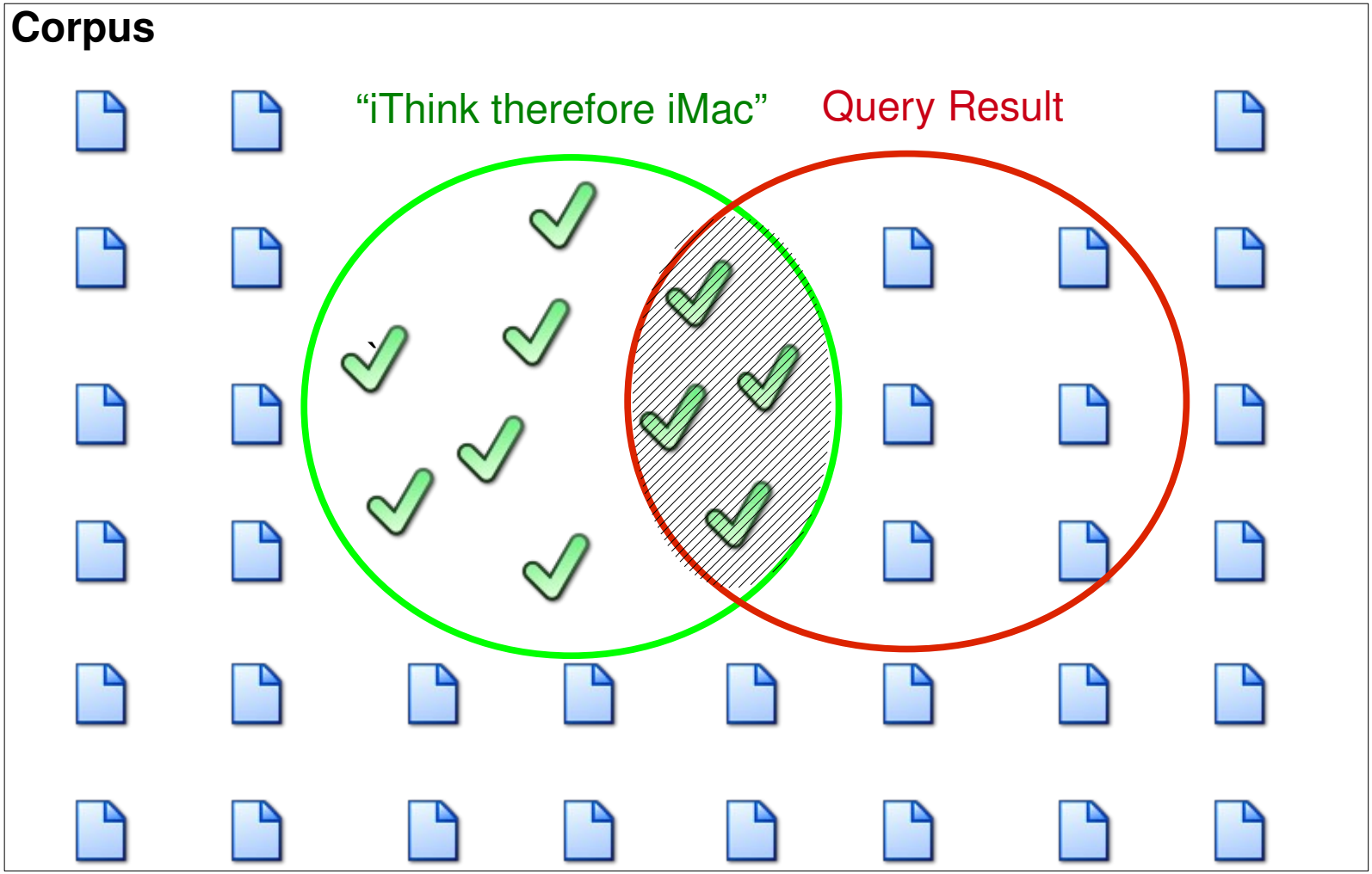
Post Processing - Classify

- Linear Regression



Merge	Group	*Classify*	

Ranking (Interestingness)



Merge	Group	Classify	Rank

Regression score

- Idea: use the value given by the decision rule
- Increase number of classes (probabilistic retrieval)

Merge	Group	*Classify*	*Rank*

Phrase Significance

Given phrase “AB”, it is significant if -

$$P(AB) > \tau$$
$$P(AB) > P(A).P(B)$$
$$P(AB) \geqslant \frac{1}{z} P(A)$$
$$P(AB) \geqslant y.P(ABC)$$

τ, y, z are parameters

Interestingness – other definitions

- Point-wise Mutual Information (PMI) based
- Weighted frequencies

Results so far ..

- Query: “Paris Hilton”
- Input phrases: 100
- Output (filtered) phrases: 30

Results so far ..

- ✓ hilton barron
- ✓ sister nicky
- hilton shares closed
- ✓ hilton 26
- paris hilton below
- ✓ hilton's shares
- ✓ stars are blind
- ✓ alcohol related reckless driving
- ✓ hilton plans
- and lodging company
- hilton appeared
- ✓ bollenbach hilton's chief executive
- ✓ the gaming and lodging
- lead hilton
- and gaming concern
- ✓ company's chairman barron hilton
- ✓ on gambling
- ✓ hilton's stock
- ✓ the conrad n. hilton foundation
- ✓ paris latsis
- ✓ the simple life 2
- ✓ acquisition of bally
- hilton owns
- ✓ the hilton brand
- ✓ paris and nicky
- hilton name
- hilton announced
- ✓ series the simple
- ✓ the newark airport hilton
- hilton lost

✓ :Interesting Phrases

Roadmap

- Approximate grouping using POS Tags
- Quality study
- User study
- Reverse indexing
- POS-tagging down to query processing
- Phrase clustering in forward index

Thanks!