Approaches for Personalized Text Summarization

Dr. Mourad Oussalah

Electronics, Electrical and Computer Engineering, University of Birmingham

Birmingham BTG Research Away Day, 20-21st September, 2010

Background

What happened?

MILAN, Italy, April 18. **A small airplane crashed** into a government building in heart of Milan, **setting the top floors on fire, Italian police reported**. There were **no immediate reports on casualties** as rescue workers attempted to clear the area in the city's financial district. **Few de** ailab about it immediately set off fears that it **might be a terrorist act** akin to the Sept. 11 attacks in the United States. Those fears sent to session lows in late morning trading.

When, where?

How many victims?

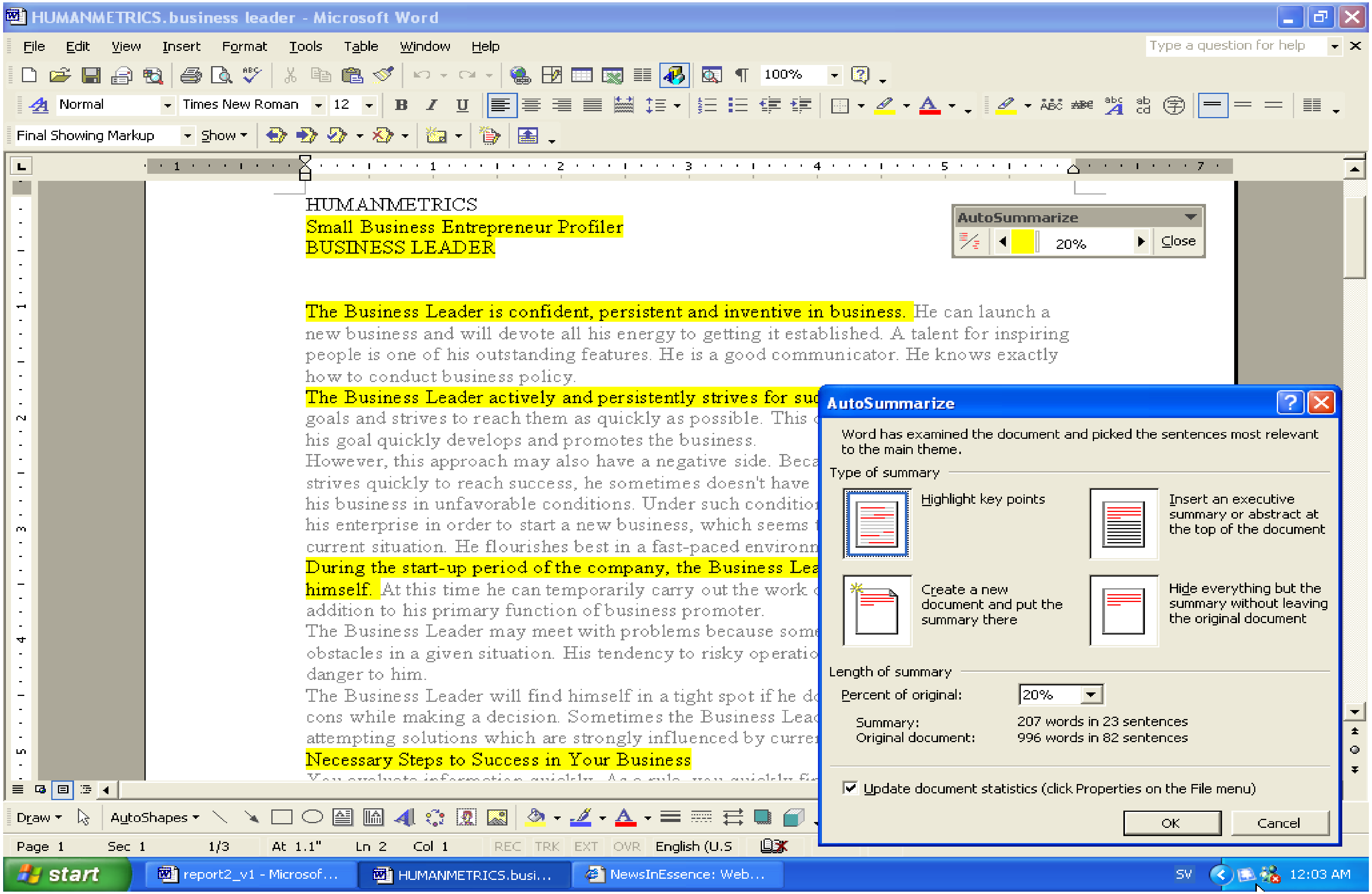
Says who?

Was it a terrorist act?

Witnesses reported hearing a loud explosion fr office building, **which houses the administrative offices of the local** Lombardy region and sits next to the city's central train station. **Italian state television** said the crash put **a hole in the 25th floor of the Pirelli building**. News reports said smoke poured from the opening. Police and ambulances rushed to the building in downtown Milan. **No further details were immediately available**.

What was the target?

MSWord AutoSummarize



HUMANMETRICS
Small Business Entrepreneur Profiler
BUSINESS LEADER

The Business Leader is confident, persistent and inventive in business. He can launch a new business and will devote all his energy to getting it established. A talent for inspiring people is one of his outstanding features. He is a good communicator. He knows exactly how to conduct business policy.

The Business Leader actively and persistently strives for suc
goals and strives to reach them as quickly as possible. This
his goal quickly develops and promotes the business.
However, this approach may also have a negative side. Beca
strives quickly to reach success, he sometimes doesn't have
his business in unfavorable conditions. Under such conditio
his enterprise in order to start a new business, which seems
current situation. He flourishes best in a fast-paced environ
During the start-up period of the company, the Business Lea
himself. At this time he can temporarily carry out the work
addition to his primary function of business promoter.
The Business Leader may meet with problems because some
obstacles in a given situation. His tendency to risky operatio
danger to him.
The Business Leader will find himself in a tight spot if he do
cons while making a decision. Sometimes the Business Lead
attempting solutions which are strongly influenced by curre
Necessary Steps to Success in Your Business

AutoSummarize

Word has examined the document and picked the sentences most relevant to the main theme.

Type of summary

- Highlight key points
- Insert an executive summary or abstract at the top of the document
- Create a new document and put the summary there
- Hide everything but the summary without leaving the original document

Length of summary

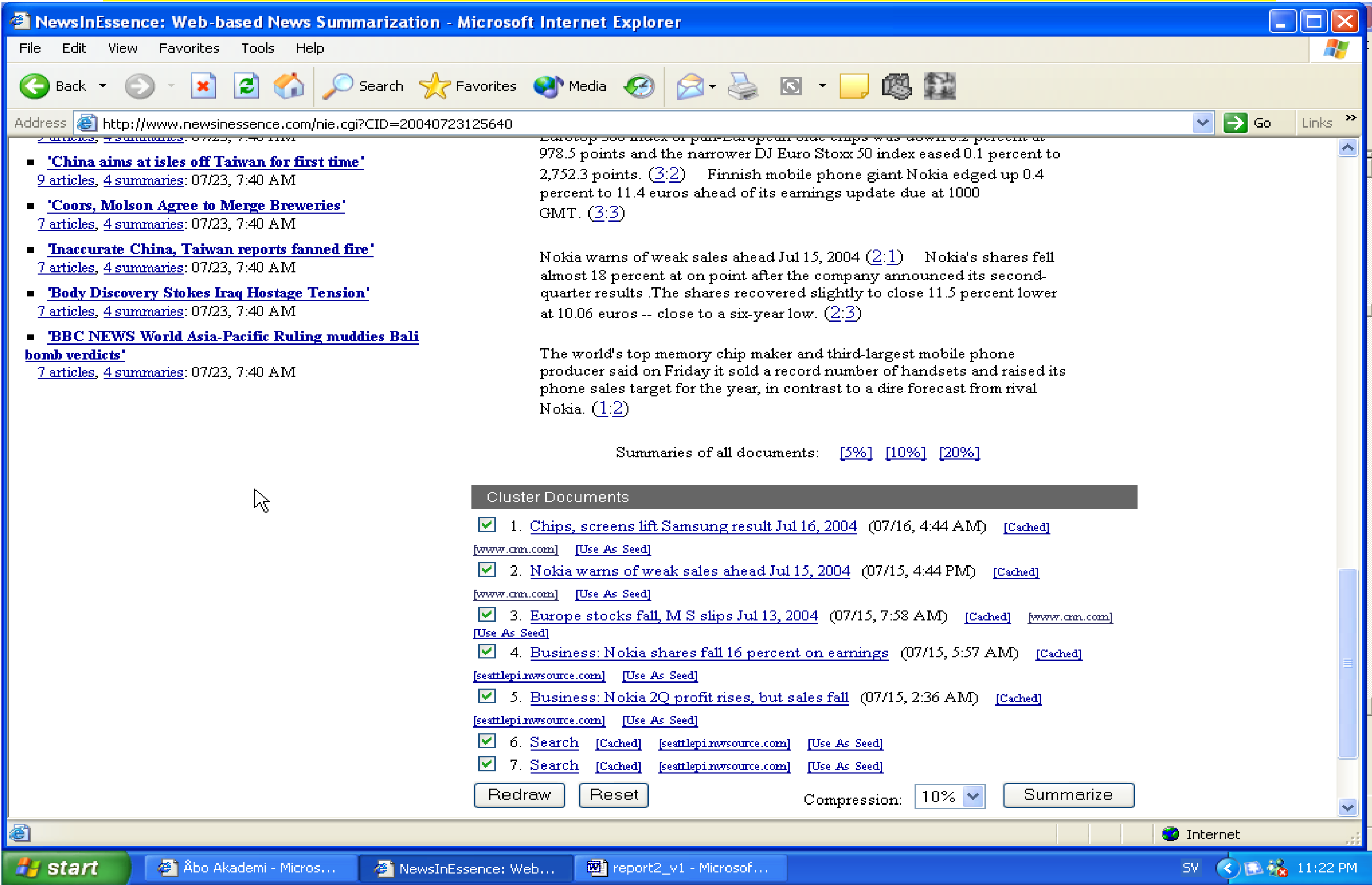
Percent of original: 20%

Summary: 207 words in 23 sentences
Original document: 996 words in 82 sentences

☑ Update document statistics (click Properties on the File menu)

OK Cancel

MEAD/NewsINEssence (Radev et al, 2003)



9 articles, 4 summaries: 07/23, 7:40 AM

- 'China aims at isles off Taiwan for first time'
  9 articles, 4 summaries: 07/23, 7:40 AM
- 'Coors, Molson Agree to Merge Breweries'
  7 articles, 4 summaries: 07/23, 7:40 AM
- 'Inaccurate China, Taiwan reports fanned fire'
  7 articles, 4 summaries: 07/23, 7:40 AM
- 'Body Discovery Stokes Iraq Hostage Tension'
  7 articles, 4 summaries: 07/23, 7:40 AM
- 'BBC NEWS World Asia-Pacific Ruling muddies Bali bomb verdicts'
  7 articles, 4 summaries: 07/23, 7:40 AM

Eurotop 300 index of pan-European blue chips was down 0.2 percent at 978.5 points and the narrower DJ Euro Stoxx 50 index eased 0.1 percent to 2,752.3 points. (3:2) Finnish mobile phone giant Nokia edged up 0.4 percent to 11.4 euros ahead of its earnings update due at 1000 GMT. (3:3)

Nokia warns of weak sales ahead Jul 15, 2004 (2:1) Nokia's shares fell almost 18 percent at on point after the company announced its second-quarter results .The shares recovered slightly to close 11.5 percent lower at 10.06 euros -- close to a six-year low. (2:3)

The world's top memory chip maker and third-largest mobile phone producer said on Friday it sold a record number of handsets and raised its phone sales target for the year, in contrast to a dire forecast from rival Nokia. (1:2)

Summaries of all documents: [5%] [10%] [20%]

Cluster Documents

☑ 1. Chips, screens lift Samsung result Jul 16, 2004 (07/16, 4:44 AM) [Cached] [www.cnn.com] [Use As Seed]
☑ 2. Nokia warns of weak sales ahead Jul 15, 2004 (07/15, 4:44 PM) [Cached] [www.cnn.com] [Use As Seed]
☑ 3. Europe stocks fall, M S slips Jul 13, 2004 (07/15, 7:58 AM) [Cached] [www.cnn.com] [Use As Seed]
☑ 4. Business: Nokia shares fall 16 percent on earnings (07/15, 5:57 AM) [Cached] [seattlepi.nwsource.com] [Use As Seed]
☑ 5. Business: Nokia 2Q profit rises, but sales fall (07/15, 2:36 AM) [Cached] [seattlepi.nwsource.com] [Use As Seed]
☑ 6. Search [Cached] [seattlepi.nwsource.com] [Use As Seed]
☑ 7. Search [Cached] [seattlepi.nwsource.com] [Use As Seed]

Redraw Reset Compression: 10% Summarize

Text summarization

- Key issues:
  - *how to identify the most important content out of the rest of the text?*
  - *how to synthesize the substance and formulate a summary text based on the identified content?*
  - How to account for semantic aspect?
- Major approaches:
  - **Selection based**: produce ”extracts”
  - Text **understanding** based: produce ”abstracts”

Types of Summaries

- Purpose:

  - Indicative, Informative, and Critical

- Form:

  - Extracts [key paragraphs, sentences, phrase] → Highly dominant
  - Abstracts (a concise summary of the central subject matter of a document” [Paice90].)

- Dimensions:

  - Single-Document, and multi-document
  - Query-dependent vs query independent

- Personalization

-via guided queries

-via specialized ontology

Approach for extractive summarization task

→ Based on the use of principle of scoring sentences.   This takes into account:

- Occurrence of Named Entity / Context
- Semantic similarity
- Positioning /title
- Redundancy / diversity
- Weighted aggregation

Features Used for Sentences Scoring

❖Named Entities

- **Persons:** Director Eugenio Cabral, Gilbert, Debby
- **Organizations:** National Hurricane Center, National Weather Center
- **Locations:** Puerto Rico, eastern Caribbean , Miami, Barahona, San Juan

❖Semantic Similarity

- Computed with the aid of **WordNet** using two large sets of previously computed similarity matrices between a large number of nouns and verbs
- Compute semantic similarity between **Title/Query** and each sentence
- Compute Semantic similarity between **each sentence** and other sentences

❖Sentences Location

Score Computation

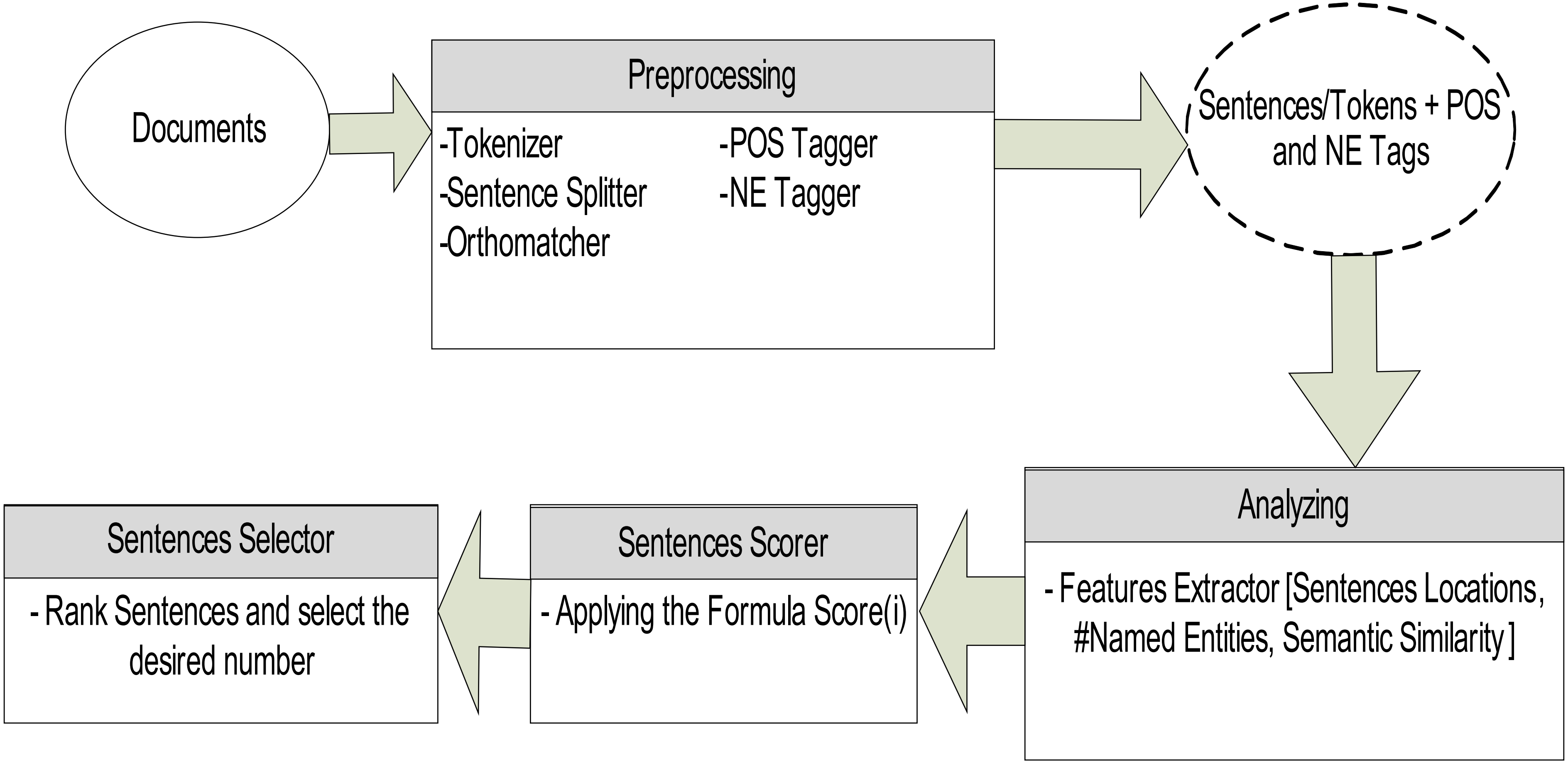
Method (1)

$$\text{Score}(i) = \frac{(\alpha\ \text{Sim}(si\ ,T) + \beta\ \text{Sim}(si\ ,Q))\ \ n(si)\ \ (\text{FNE}\ (si) + 1)\ \ P(si)}{N\ \ (\text{NE} + 1)}$$

Where:

- Score(i) is the score of sentence (i)
- N = the total number of sentences
- (α + β = 1).
- n(si) = The number of sentences that have semantic similarity score bigger than a pre-defined threshold value
- P(s) = either 1 for sentences appearing at the top and end of the document, or 0.5 for the rest.
- Sim(si ,T) and Sim(si ,Q) are for the Semantic Similarity between the Title and the Query, respectively, and the sentence (i).
- FNE (si) = the number of Named Entities contained in the sentence (i)
- NE: the number of Named Entities in the document.

Architecture of the Developed System



Documents

Preprocessing

-Tokenizer

-Sentence Splitter

-Orthomatcher

-POS Tagger

-NE Tagger

Sentences/Tokens + POS and NE Tags

Analyzing

- Features Extractor [Sentences Locations, #Named Entities, Semantic Similarity]

Sentences Scorer

- Applying the Formula Score(i)

Sentences Selector

- Rank Sentences and select the desired number

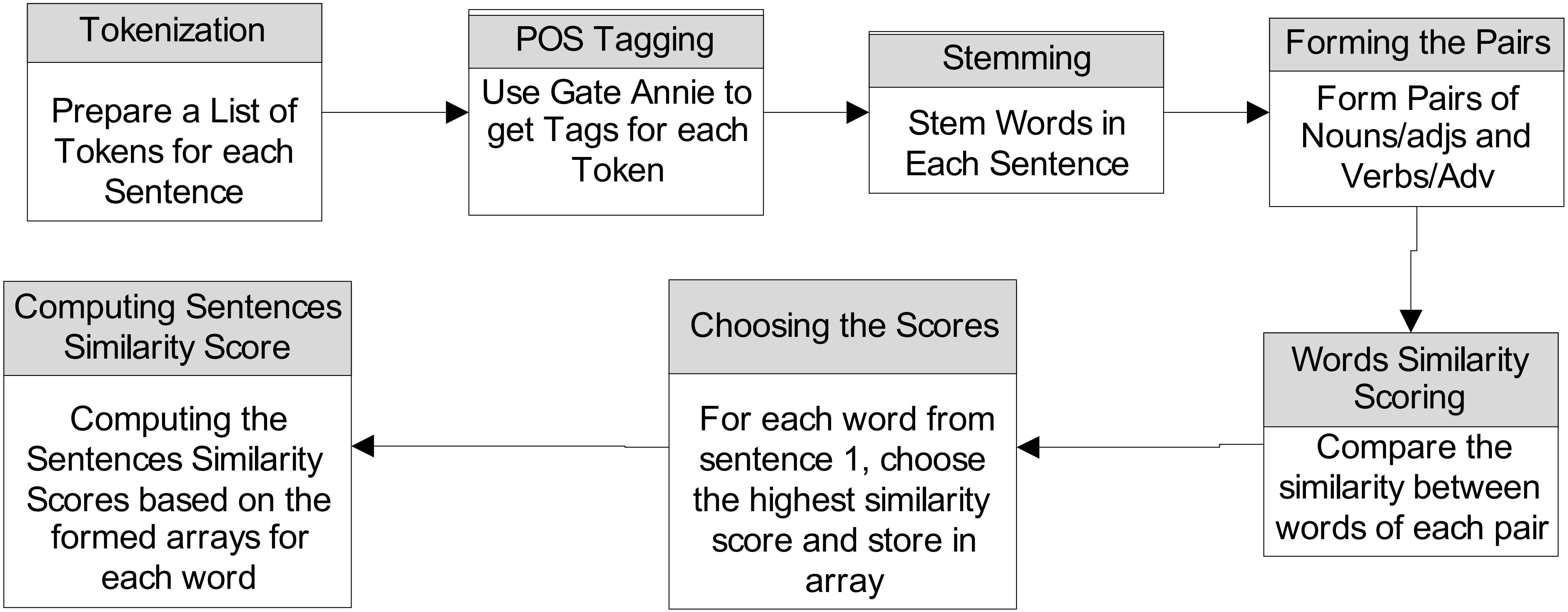
Similarity Between Sentences

Average over all words of sentence

$$\text{idf-modified-cosine}(x,y) = \frac{\sum_{w \in x,y} \text{tf}_{w,x}\text{tf}_{w,y}(\text{idf}_w)^2}{\sqrt{\sum_{x_i \in x}(\text{tf}_{x_i,x}\text{idf}_{x_i})^2} \times \sqrt{\sum_{y_i \in y}(\text{tf}_{y_i,y}\text{idf}_{y_i})^2}}$$

But, not very effective
- adverbs, adjectives not handled

Other approaches
-Use WordNet to extract nouns associated to each word in sentences and perform above expression **, or**
- Restrict to **highest** pair similarity value



Tokenization

Prepare a List of Tokens for each Sentence

POS Tagging

Use Gate Annie to get Tags for each Token

Stemming

Stem Words in Each Sentence

Forming the Pairs

Form Pairs of Nouns/adjs and Verbs/Adv

Words Similarity Scoring

Compare the similarity between words of each pair

Choosing the Scores

For each word from sentence 1, choose the highest similarity score and store in array

Computing Sentences Similarity Score

Computing the Sentences Similarity Scores based on the formed arrays for each word

Method 2 Use of Redundancy/Diversity

Idea: Reduce redundancy and increase the diversity

- **Redundancy:**
- **Average Semantic Similarity between two Sentences**

$$R = \frac{|s1 \cap s2|}{Max(|s_1|,|s_2|)}$$

- **Two metrics:**
  - simple words matching
  - semantic similarity exceeding a threshold

- **Diversity:**
  - **Two Metrics:**
    - **With the usage of Antonyms**
    - **Without**

$$D_{Ant} = \frac{1}{N}\sum_i sim(w_{1i}, Ant(w_{2i}))$$

$$D_{Ant} = \frac{1}{N}\sum_i sim(w_{2i}, Ant(w_{1i}))$$

Method 2 Use of Redundancy/Diversity

Score of sentence (i) = min $_j$ [R(i,j) – D(i,j)] * a *b

a and b account for location and similarity with respect to title/query

Alternative

MMR (maximal Marginal relevance)

$$MMR = Arg \max_{D_i \in R \backslash S} \left[\lambda . Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right]$$

Method 3
Use of Wikipedia

- Instead of use of wordNet semantic similarity, the page rank like based approach is approached.
- Use Wikipedia
- E,g., similarity between (cat, animal) is constructed by looking at number of documents where both cat and animal occur together, up to a normalization factor

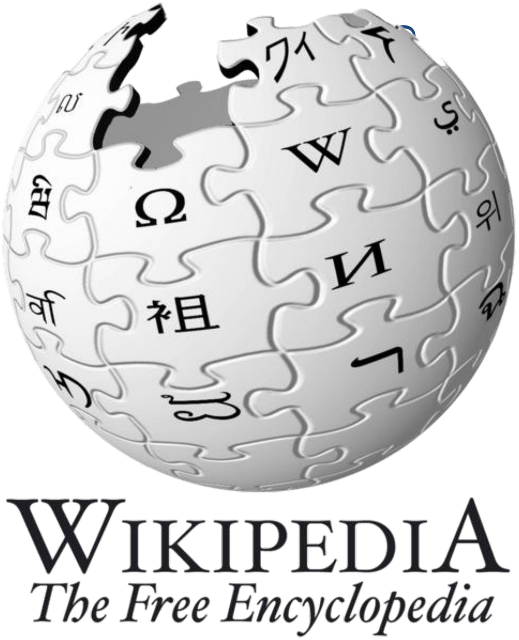
Background

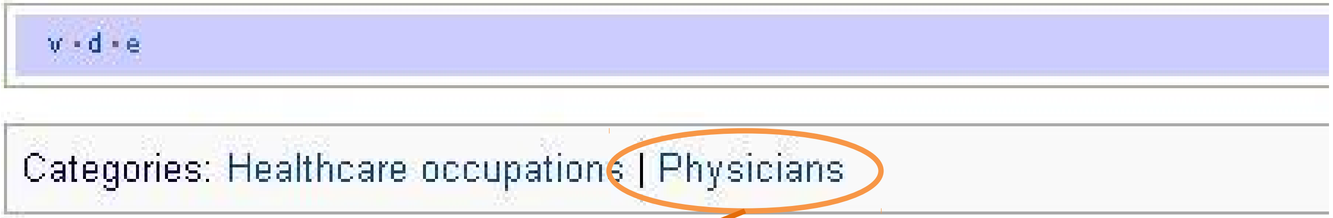
- Wikipedia is the **largest known encyclopedia** to date
  - English version has over 3.3 million articles and 600 million words

- Each article discusses a **single unique subject**
  - we use the article title to represent the ***concept*** discussed between articles

- **Hierarchical Categories** exist to organize articles
  - Each article belongs to at least one ***category***

- Our approach relies on the ***information and the structure*** of Wikipedia to compute the ***relatedness*** between concepts and use it in the task of WSD

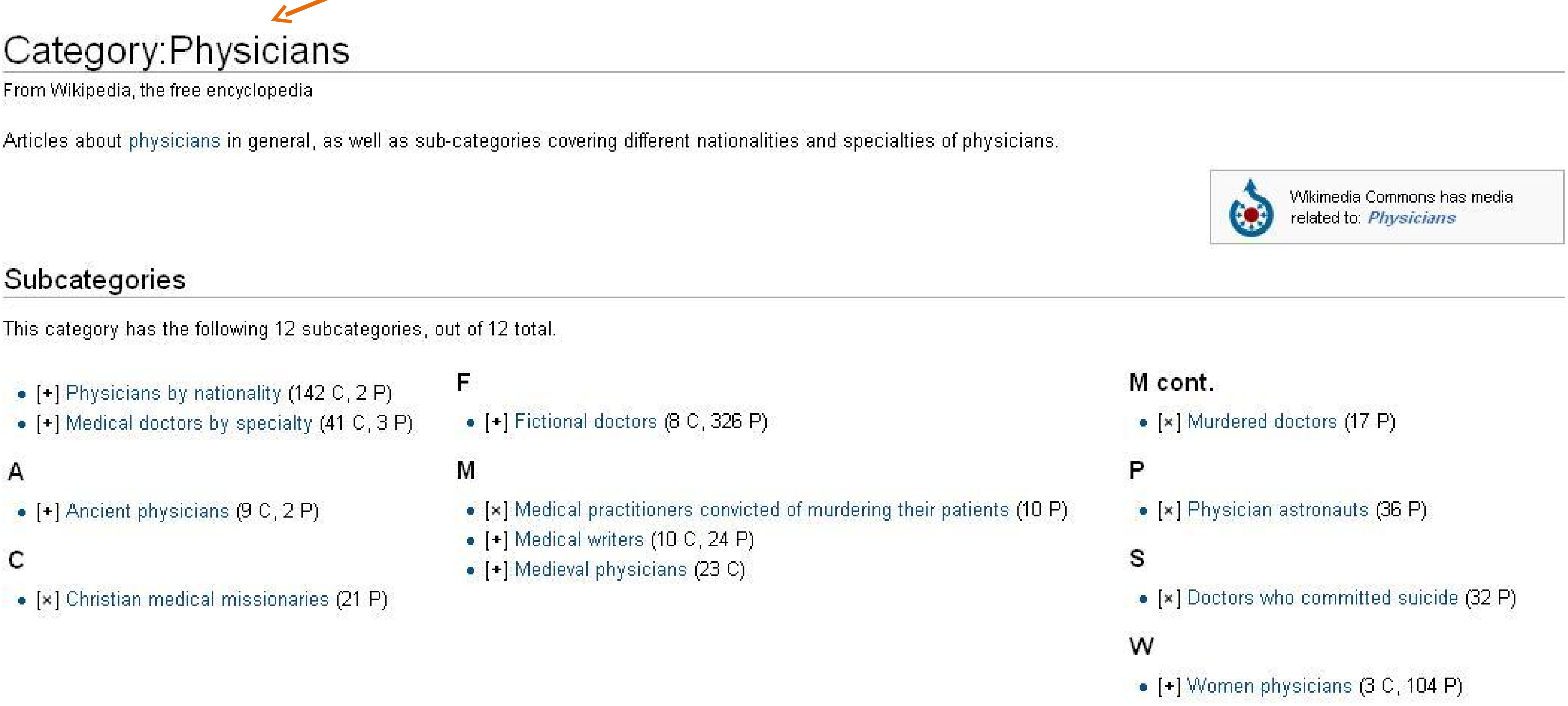




v • d • e

Categories: Healthcare occupations | Physicians

The concept **Physician** belongs to two categories: **Physicians and Healthcare Occupations**



Category:Physicians

From Wikipedia, the free encyclopedia

Articles about physicians in general, as well as sub-categories covering different nationalities and specialties of physicians.

Wikimedia Commons has media related to: *Physicians*

Subcategories

This category has the following 12 subcategories, out of 12 total.

- [+] Physicians by nationality (142 C, 2 P)
- [+] Medical doctors by specialty (41 C, 3 P)

A

- [+] Ancient physicians (9 C, 2 P)

C

- [×] Christian medical missionaries (21 P)

F

- [+] Fictional doctors (8 C, 326 P)

M

- [×] Medical practitioners convicted of murdering their patients (10 P)
- [+] Medical writers (10 C, 24 P)
- [+] Medieval physicians (23 C)

M cont.

- [×] Murdered doctors (17 P)

P

- [×] Physician astronauts (36 P)

S

- [×] Doctors who committed suicide (32 P)

W

- [+] Women physicians (3 C, 104 P)

Term-Concepts Table

- The weight of each term in an article is computed
  - We use the TFIDF weight measure
- For a term $\boldsymbol{t_i}$, its weight $\boldsymbol{w_i}$ in an article ***c*** resembles its association strength with the article ***c***
- For each term, a vector of its weights in all the Wikipedia articles is constructed. The larger the weight, the more related the term is to the article
- After constructing the vectors for each term, we apply a ***boosting*** algorithm. The purpose of this algorithm is twofold:
  - Handling the occurrence of some important terms in the redirect links but not in the content of the articles
  - Increasing the importance level of the articles containing key terms in their titles
  - Increasing importance level of articles containing words of ontology

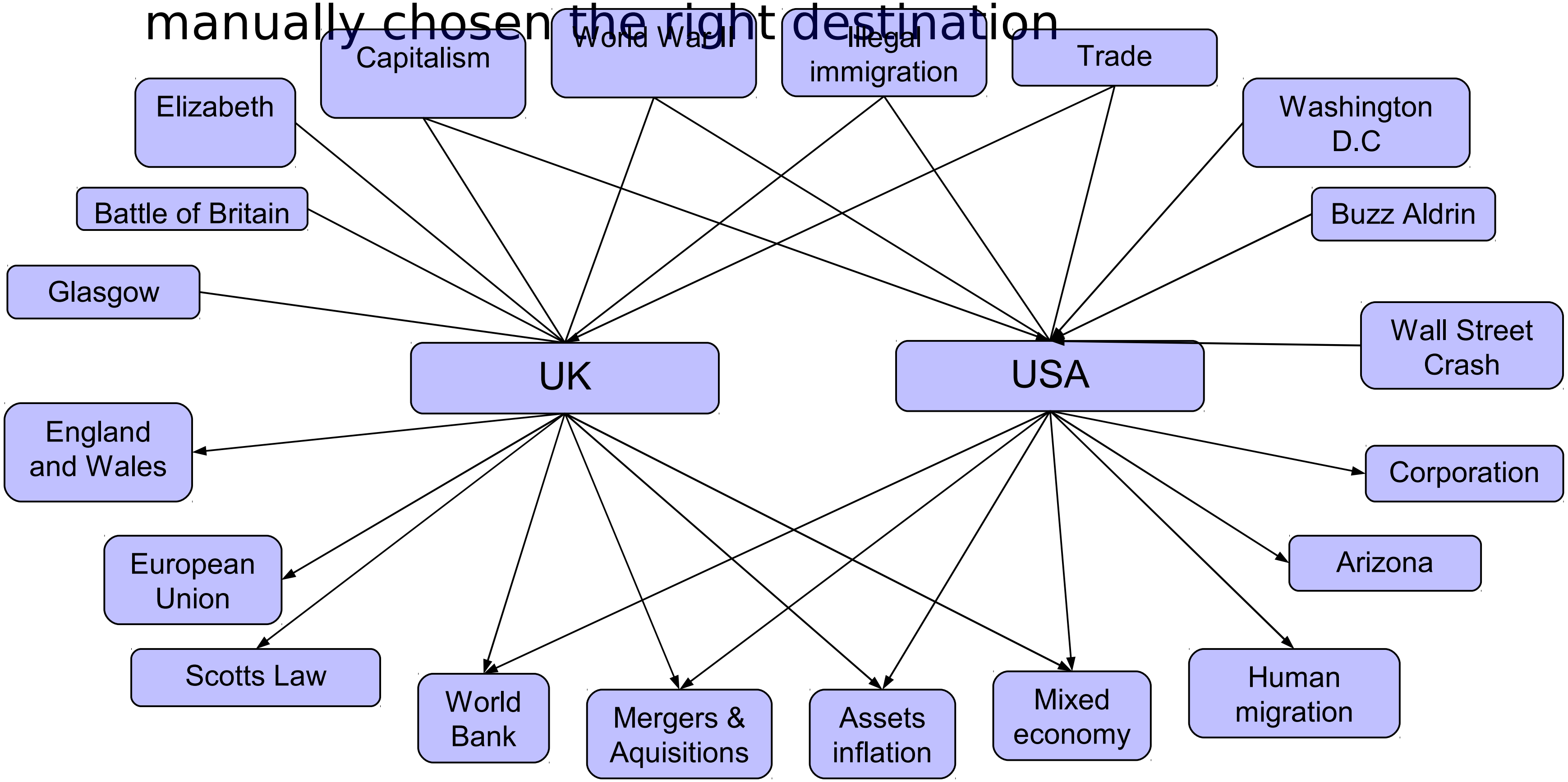
Example 1

- For the terms ***Unhappy*** and ***Jobless***, the following lists of most related concepts were built

	Unhappy	***Unhappy* (Boosted)**	***Jobless***	***Jobless* (Boosted)**
1	Implications of Divorce	Depression (mood)	Growth Recession	Unemployment
2	Unhappy Consciousness	Unhappy Consciousness	When Work Disappears	Jobless Recovery
3	The Better Half	Implications of Divorce	Pôle Emploi	James Renshaw Cox
4	The Human Contract	Unhappy Triad	James Renshaw Cox	Growth Recession
5	Kurumi Enomoto	Fan the Flame	Joe Ma Wai-ho	When Work Disappears
6	Pamela Springsteen	Unhappy Happiness	Vetti	Pôle Emploi
7	Tristan Davies	Happy Number	Volksgrenadier	Joe Ma Wai-ho
8	Fan the Flame	the Better Half	shadowstats.com	Vetti
9	Notes & Rhymes	the Human Contract	Jobless Recovery	Volksgrenadier
10	Ballad of a Teenage Queen	Kurumi Enomoto	Imperfect Competition	shadowstats.com

Wikipedia Links

- Links between Wikipedia articles provide the reader the chance to explore other related articles while reading one
- For every link in Wikipedia, a human editor has manually chosen the right destination



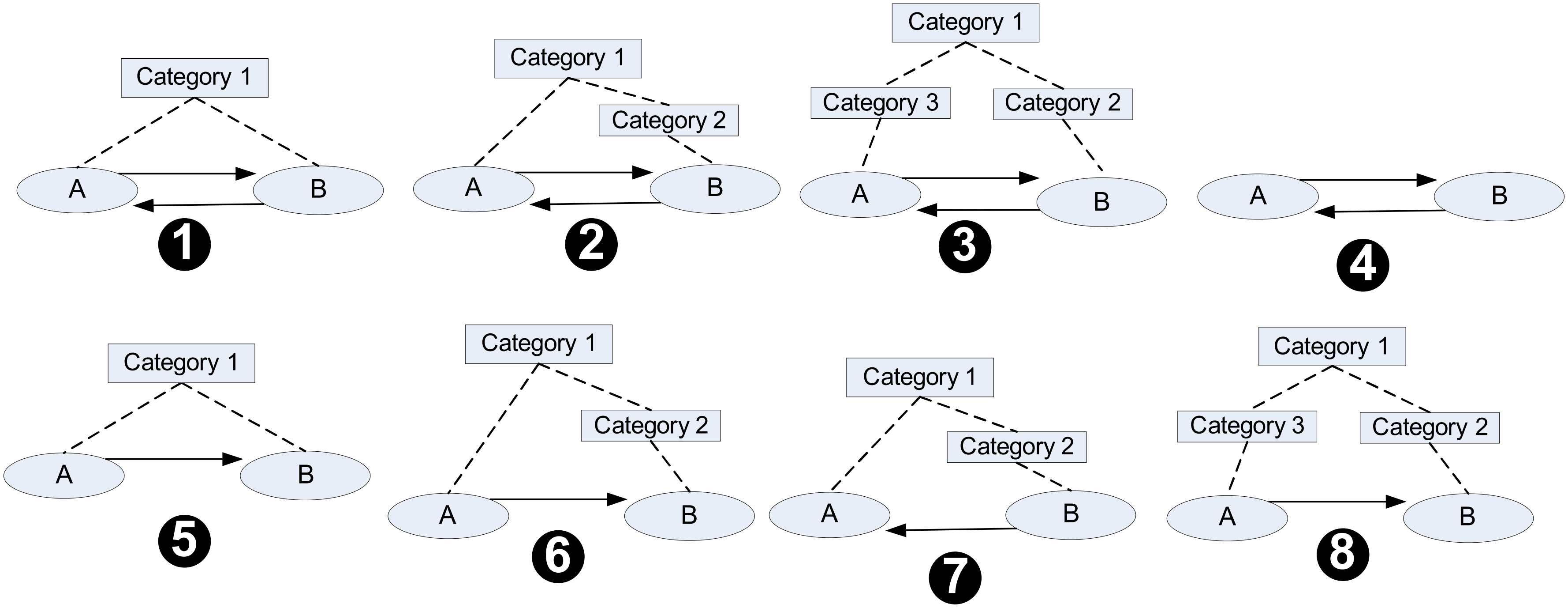
Wikipedia Links and Categories Structure

- Not all links are of the same importance
  - e.g. ***Peripheral Vision*** and ***Basketball Court*** are links existing within the ***Basketball*** article

- Some articles have very large number of links
  - E.g. ***UK*** have over 70,000 incoming links

- Therefore, links classification is applied by utilizing the following:
  - Link type (internal, first passage, '*See Also*')
  - Link direction (incoming or outgoing)
  - Number of links shared between two articles
  - Categories shared between articles

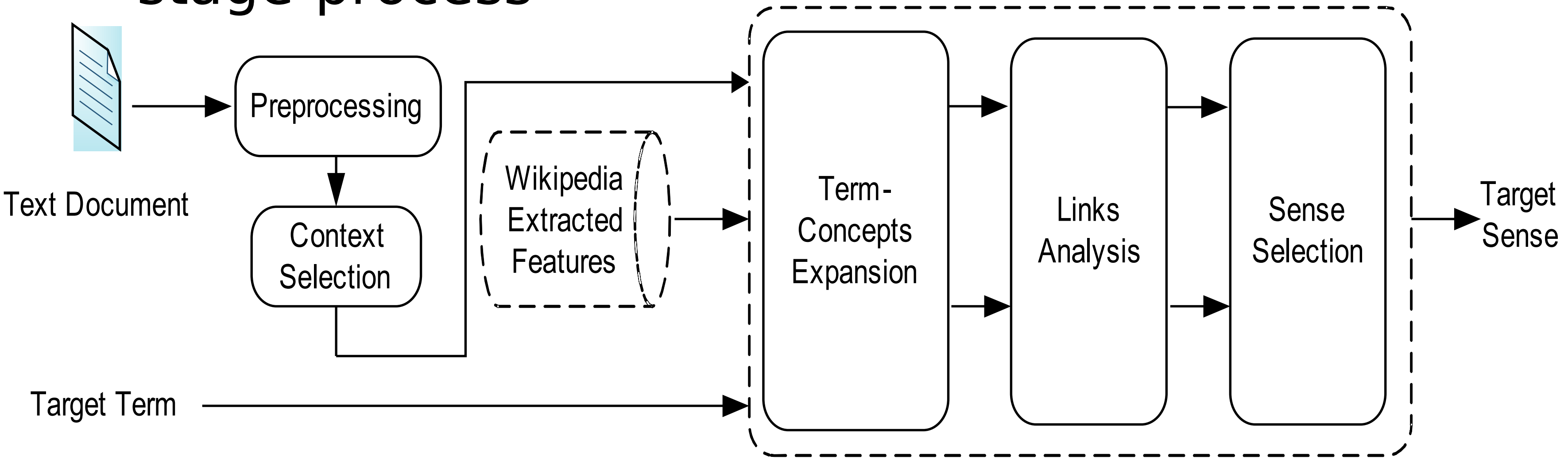
Wikipedia Links and Categories Structure



Application 1. Word Sense Disambiguation

- Determine the right sense of a term based on the **context** it appears in

- The previously-extracted features from Wikipedia are used for the task in a two stage-process



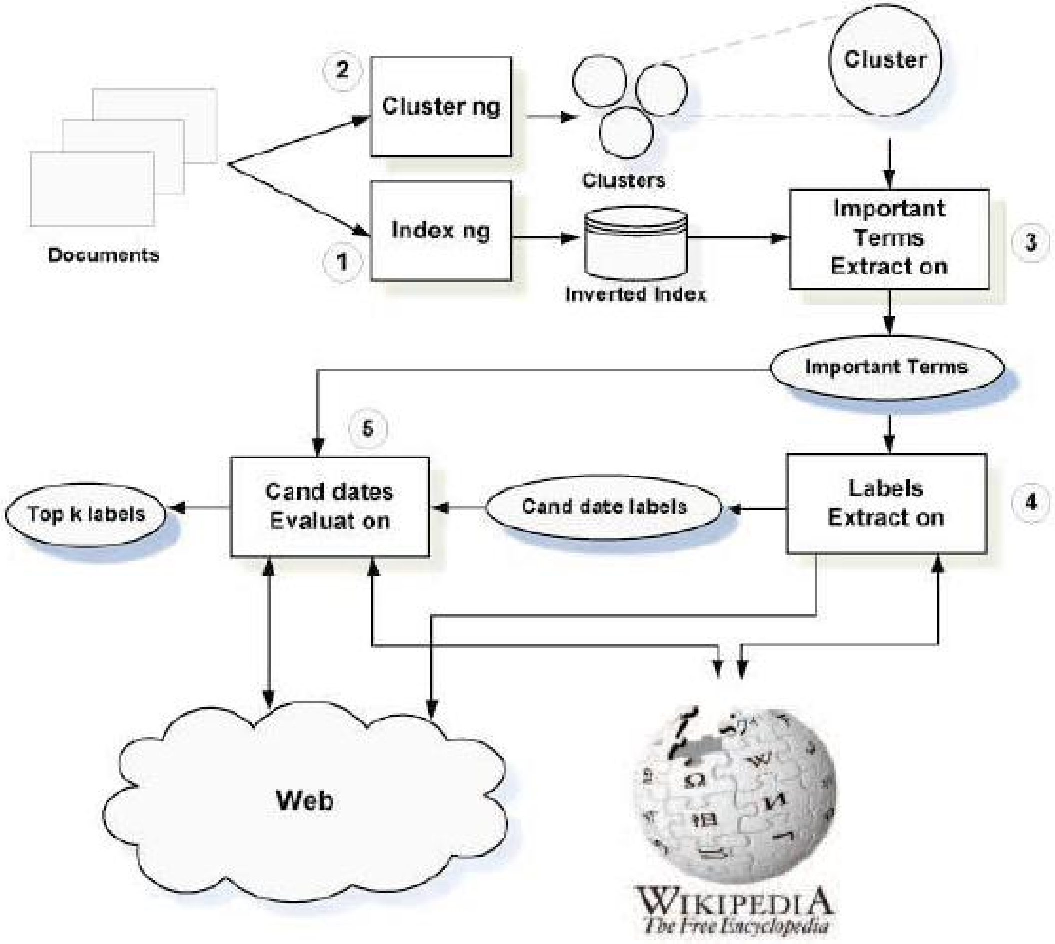
Application 2. Clusters Labeling

- Use of concepts titles to represent clusters

- Finding the most suitable concepts based on examining the dominant concepts within each cluster

- Generate a list of possible Candidate Labels

- Evaluate Candidate Labels and choose the best after keywords-boosting



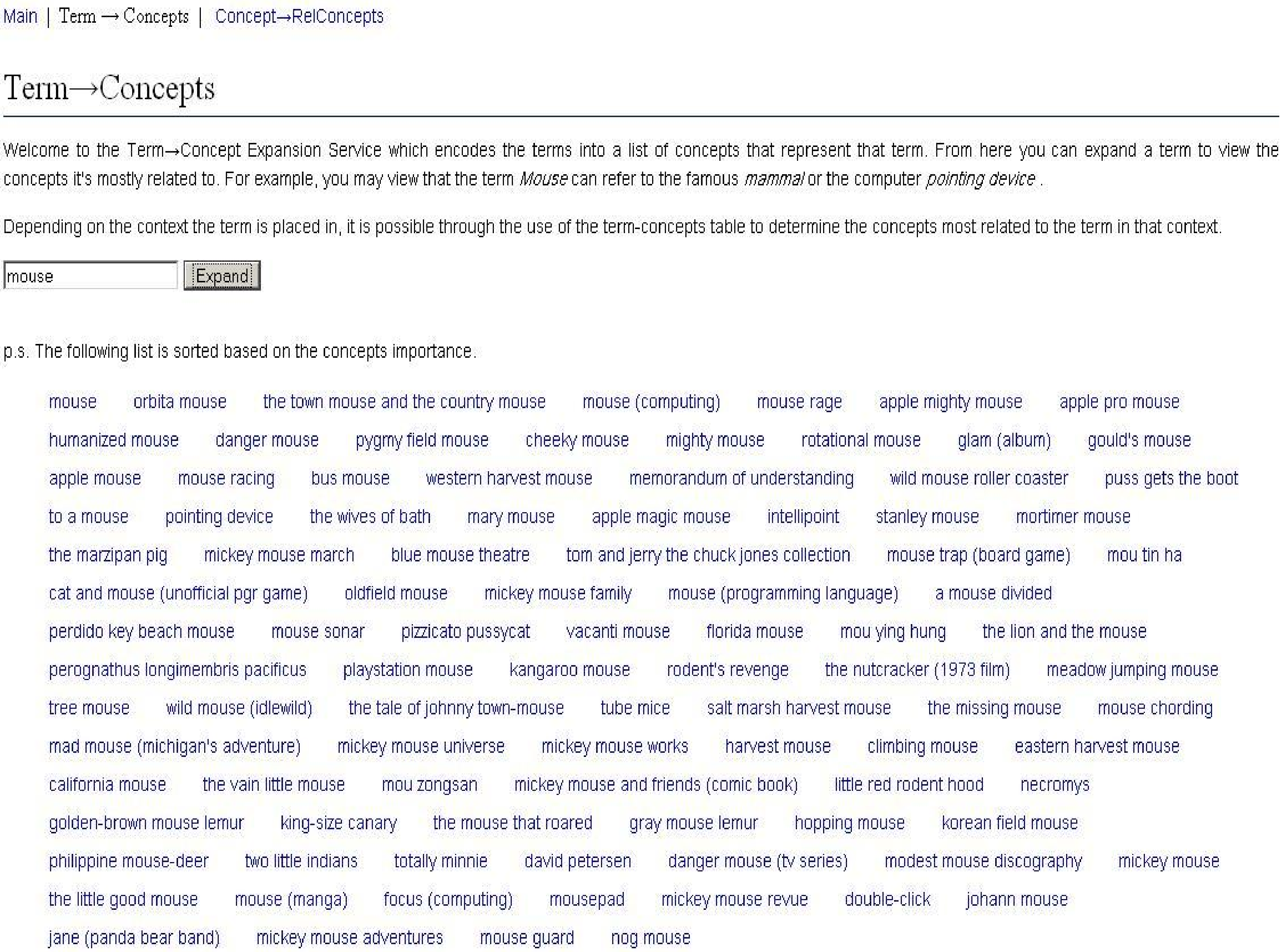
A general framework for Clusters Labeling*

* D. Carmel, H. Roitman, and N. Zwerdling. 2009. Enhancing Cluster Labeling using Wikipedia. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pages 139–146. ACM.

Application 3. Extracting Content Holes Within Documents

- Helps view the content of a document from **multiple perspectives** by presenting strongly ***related*** but ***different*** concepts from those existing within a document

- Searches the document for missing information (holes) and present them to the user



Main | Term → Concepts | Concept→RelConcepts

Term→Concepts

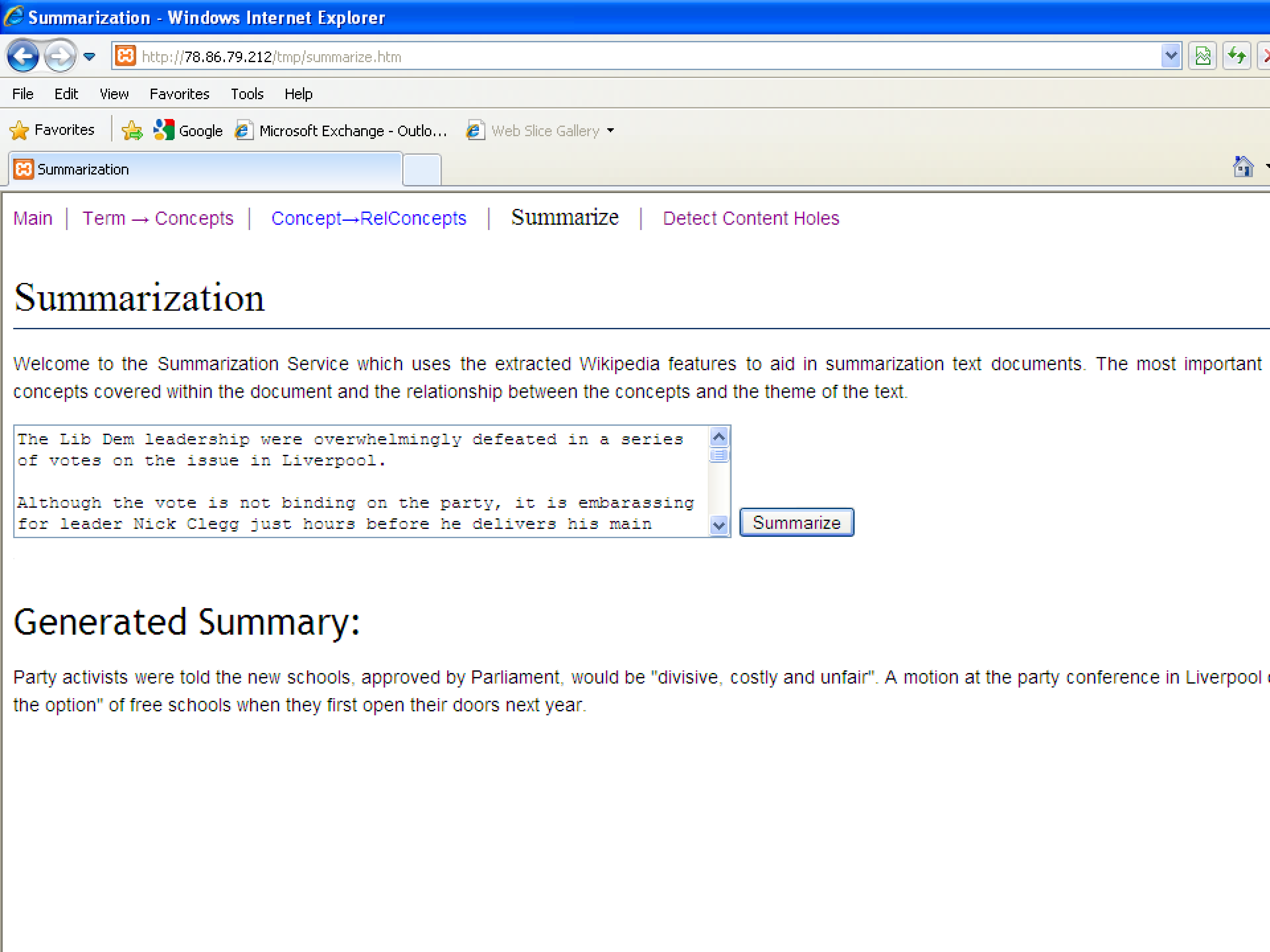
Welcome to the Term→Concept Expansion Service which encodes the terms into a list of concepts that represent that term. From here you can expand a term to view the concepts it's mostly related to. For example, you may view that the term *Mouse* can refer to the famous *mammal* or the computer *pointing device* .

Depending on the context the term is placed in, it is possible through the use of the term-concepts table to determine the concepts most related to the term in that context.

mouse [Expand]

p.s. The following list is sorted based on the concepts importance.

mouse orbita mouse the town mouse and the country mouse mouse (computing) mouse rage apple mighty mouse apple pro mouse
humanized mouse danger mouse pygmy field mouse cheeky mouse mighty mouse rotational mouse glam (album) gould's mouse
apple mouse mouse racing bus mouse western harvest mouse memorandum of understanding wild mouse roller coaster puss gets the boot
to a mouse pointing device the wives of bath mary mouse apple magic mouse intellipoint stanley mouse mortimer mouse
the marzipan pig mickey mouse march blue mouse theatre tom and jerry the chuck jones collection mouse trap (board game) mou tin ha
cat and mouse (unofficial pgr game) oldfield mouse mickey mouse family mouse (programming language) a mouse divided
perdido key beach mouse mouse sonar pizzicato pussycat vacanti mouse florida mouse mou ying hung the lion and the mouse
perognathus longimembris pacificus playstation mouse kangaroo mouse rodent's revenge the nutcracker (1973 film) meadow jumping mouse
tree mouse wild mouse (idlewild) the tale of johnny town-mouse tube mice salt marsh harvest mouse the missing mouse mouse chording
mad mouse (michigan's adventure) mickey mouse universe mickey mouse works harvest mouse climbing mouse eastern harvest mouse
california mouse the vain little mouse mou zongsan mickey mouse and friends (comic book) little red rodent hood necromys
golden-brown mouse lemur king-size canary the mouse that roared gray mouse lemur hopping mouse korean field mouse
philippine mouse-deer two little indians totally minnie david petersen danger mouse (tv series) modest mouse discography mickey mouse
the little good mouse mouse (manga) focus (computing) mousepad mickey mouse revue double-click johann mouse
jane (panda bear band) mickey mouse adventures mouse guard nog mouse


Summarization - Windows Internet Explorer
http://78.86.79.212/tmp/summarize.htm
File Edit View Favorites Tools Help
Favorites
Google
Microsoft Exchange - Outlo...
Web Slice Gallery
Summarization
Main | Term → Concepts | Concept→RelConcepts | Summarize | Detect Content Holes
Summarize

Summarization

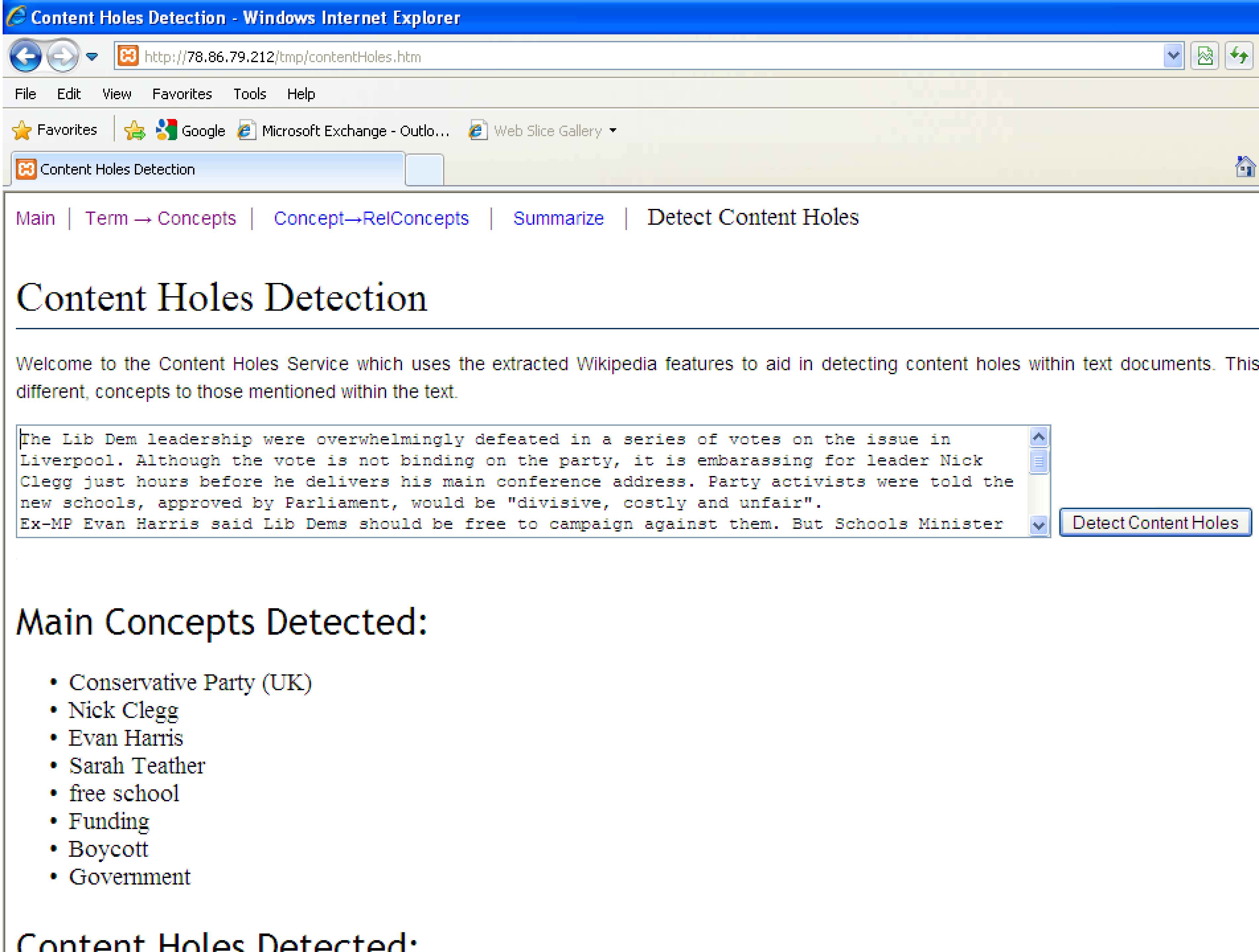
Welcome to the Summarization Service which uses the extracted Wikipedia features to aid in summarization text documents. The most important concepts covered within the document and the relationship between the concepts and the theme of the text.

```
The Lib Dem leadership were overwhelmingly defeated in a series
of votes on the issue in Liverpool.

Although the vote is not binding on the party, it is embarassing
for leader Nick Clegg just hours before he delivers his main
```

Generated Summary:

Party activists were told the new schools, approved by Parliament, would be "divisive, costly and unfair". A motion at the party conference in Liverpool the option" of free schools when they first open their doors next year.



Content Holes Detection

Welcome to the Content Holes Service which uses the extracted Wikipedia features to aid in detecting content holes within text documents. This
different, concepts to those mentioned within the text.

```
The Lib Dem leadership were overwhelmingly defeated in a series of votes on the issue in
Liverpool. Although the vote is not binding on the party, it is embarassing for leader Nick
Clegg just hours before he delivers his main conference address. Party activists were told the
new schools, approved by Parliament, would be "divisive, costly and unfair".
Ex-MP Evan Harris said Lib Dems should be free to campaign against them. But Schools Minister
```

Main Concepts Detected:

- Conservative Party (UK)
- Nick Clegg
- Evan Harris
- Sarah Teather
- free school
- Funding
- Boycott
- Government

Content Holes Detected:

Testing: Summarization Tasks for TAC (Text Analysis Conference) 08, 09, 10, NIST

Two Tasks:

- Write a short (~ 100-word) summary of a set of newswire articles, under the assumption that the user has already read a given set of earlier articles.
- Write summaries of opinions from blogs. Questions from will be given and the text snippets output by QA systems. Required is the production of short coherent summaries of the answers to the questions, either from the text snippets themselves, or from the associated documents
- **Algorithm performed well and good**