clearyk6 / CA4009 Public

Code Issues Pull requests Actions Projects Security Insights

master

CA4009 / Notes / Chapter4.md

clearyk6 Notes on Text and Web Retrieval History

1 contributor

131 lines (84 sloc) 5.36 KB

Summarization

Snippet Creation: Creating digestable snippets of a document to allow the user to decide a documents relevance at a glance, displayed in order of relevance in the SERP

SERP: Search Engine Results Page

A **summary** is a *condensed derivative of a source text*

- content reduction or generalisation on what is important in a doc
  - **selection**: forming a summary based on a subset of topical content of the source in detail
  - **generalisation**: forming a summary which overviews the entire topical contents of the source

Need to be aware of what we want from a summarisation...

- Function of the summarisation: informing or alerting?
- Audience? General or more narrowly targeted
- Level of subject knowledged assumed of the reader
- Desired output of resulting summary

Need to also consider the **format** of source doc, scientific reports differ greatly from articles

output of the sumariser affected by the following:

- form of the input text
- purpose of the summary
- format/limitations of the output space eg word limit, screen space etc

Originally used **human summarisers** so when developing automated systems, key to reflect on the human process. Generally fit to a pre-defined set of guidelines

- desired style of output summary
- degree of reduction in the amount of text
- expected audience
- format of the output

Automatic summarisation can be divided into two classes:

- **Information Extraction & Text Synthesis**
  - information extracted from doc (but summary itself is a new doc)
  - extraction occurs using *natural language processing* methods.
  - info extraction: key info (names, places, relationships between entities, actions etc) all placed into a database
  - the new text (summarisation) is synthesised using *automatic text generation* methods

- **Sentence and/or Phrase Extracting**
  - Summary composed of a subset of the sentences/phrases in the source doc. Much shallower and easier approach
  - Score all the sentences/phrases. Use of metric(s) indicating importance
  - Use highest scoring s/ps in the summary. May need to consider the context of each to make sense (i.e the use of a vague pronoun at the start of the sentence may make no contextual sense)
  - These summaries are difficult to read, but should still give the reader a good understanding of the contents of the doc

Sentence scoring for S/P Extraction:

- Luhn's score for cluster of significant words
- Frequency of the title word(s) in a sentence/phrase
- Location of sentence within doc
- Frequency of query words in doc

Luhn's Keyword Cluster Method

- Luhn determines that two words are significantly related to each other if placed within 5 *insignificant* words of each other in a doc
  - significatnt words occur between low and high frequencies (high = stop words, low = very rare words, both ignored)
- **Procedure**: locate the first significant word in a sentence
- Locate the last significant word in that sentence **or** before there is a sequence of 5 insignificant words
- Bracket the phrase with the first and last significant words at the ends. eg "*The sentence [**scoring** process utilised **information** both from the **structural**] organisation.*"
- Calculate the significance score of the bracketed phrase

```
Calculation:

SS1 = (SW)** / TW

    SS1 = setence score
    SW = no of bracketed significant words
    TW = total number of bracketed words
```

- if two or more clusters occur in the sentence, the one with the highest score is chosen as the Sentence Score

Title Term Frequency Method

- Title of a doc usually revels major info about the contents
- Therefor, sentences containing one or more title words can be considered more significant
- Title score computed for each sentence

```
Calculation:

SS2 = TTS / TTT

    SS2 = sentence title score
    TTS = Total no of title terms in the sentence
    TTT = Total no of words in the title
```

- TTT is important here as it normalises the sentence score - otherwise title sentence score could dominate all sentence scores

Location/Header Method

- The first sentence of a document & section headings often prove significant

- These sentences can be assigned a location score to boost their profile

  ```
  Calculation

  SS3 = 1 / NS

      SS3 = sentence location score
      NS = Number of sentences in a document
  ```

- Location score inversely porportional to the length of a document

Query Bias Method

- A bias factor to score sentences containing query terms

  ```
  Calculation

  SS4 = (tq)** / nq

      SS4 =
      tq = number of query terms in a sentence
      nq = number of terms in the query
  ```

- nq acts as a normalisation factor

Combining All Scores

Final Sentence Significance Score (SSS) calculated using the methods above along with **eperimentally determined scalar constants** to control each factors influence:

```
SSS = a(SS1) + b(SS2) + c(SS3) + d(SS4)
```

- The normalisation of SS2 and 4 ensure that the combo process here is balanced between each factor
- Snippets for IR tend to favour query term presence, so usually a high score given to this factor

The **optimial length** of a summary is a *compromise* between material covered in the summary and the appropriate length of the summary

"How could we evaluate the effectiveness/robustness of Snippets in the SERP"