RCOS

Context Research : NLP Context Analysis

Q What is needed to know context?

NLP: noun, verb, adjectives

(what) noun

verb ↳ action of nouns

adjectives → sentiment

→ **Themes**

↳ noun phrases w/ contextual relevance scores.

- Extracted noun phrases, which are then scored on the relevance of potential themes through lexical chaining

Text Analytics / Text Mining

Basic steps :

1. Language Identification
2. Tokenization
3. Sentence Breaking
4. Part of Speech Tagging
5. Chunking
6. Syntax Parsing
7. Sentence chaining

2. Tokenization.

- Individual units of meaning you're operating on.
  → words, phonemes, or full sentences.
- Process of breaking text documents apart into these pieces.

TODO: Research

Factors needed to identify a programming language

* Tokens can be :

Words
punctuation
hyperlinks
apostrophes

→ Programmes : Language identification
- keywords
- semicolons
- syntax

3. Sentence Breaking

• Usually denoted by a dot •

– but are all dots a sentence break?

Ex: Dr. Anisha

⇒ not a sentence break.

Likewise : not all `\n' indicate a sentence/code line break

C++ → ; indicates a line break

python → \n indicates a line break (if last character ≠ /)

4. Part of Speech Tagging

- To figure out whether a given token represents a proper noun or a common noun, or if its a verb, an adjective, or something else
- **PoS**

  How?

5. Chunking

- A range of sentence-breaking systems that splinter a sentence into its component phrases (noun phrases, verb phrases ...)
- Assigning PoS-tagged tokens to phrases.

!! a chunk of for loop

PoS tagged statements assigned to a code-scope.

Q: Will a chunk be a function, classes, loops or a bunch of sequential statements ?

→ depends on what a token is

6. Syntax Parsing (sentence diagramming)

- A way to determine the structure of a sentence
- Preparatory step in sentimental analysis

Sentimental analysis ?

Can we use ?

7. Sentence Chaining (sentence relation)

- lexical chaining

  ↳ to connect related sentences.

- Links individual sentences by each sentences' strength of association to an overall topic