3. Text Mininig

Task

- Find any suitable textual data for processing which will contain at least 500 sentences.
  - you can manually collect texts from BBC/CNN/New York Times, or
  - use the crawler from the first homework/tutorial and extend it to crawl particular website and collect content for this task, or
  - use any other suitable texts
- Perform following NLP tasks
  - POS tagging
  - NER with entity classification (using nltk.ne_chunk)
  - NER with custom patterns
    - e.g. every match of: adjective (optional) and proper noun (singular/plural) is matched as the entity
    - see slides 31 or 38 from lecture 4 [https://edux.fit.cvut.cz/courses/MI-DDW.16/_media/lectures/lecture4.pdf] for some NLTK examples using RegexpParser or custom NER
- Implement your custom entity classification
  - For each detected entity (using both nltk.ne_chunk and custom patterns)
    - Try to find a page in the Wikipedia
    - Extract the first sentence from the summary
    - Detect category from the sentence as a noun phrase
      - Example:
        - for „Wikipedia‟ entity the first sentence is „Wikipedia (/ˌwɪkɨˈpiːdiə/ or /ˌwɪkiˈpiːdiə/ WIK-i-PEE-dee-ə) is a free online encyclopedia that aims to allow anyone to edit articles.‟
        - you can detect pattern ‟… **is**/VBZ a/DT free/JJ online/NN encyclopedia/NN …‟
        - the output can be „Wikipedia‟: „free online encyklopedia‟
    - For unknown entities assign default category e.g. „Thing‟

Wikipedia package in Python:

```
import wikipedia
results = wikipedia.search("Wikipedia")
print(results)
page = wikipedia.page("Wikipedia")
print(page.summary)
```

Instructions for submitting

In your private namespace on EDUX provide the following information:

- Description of the data you used for processing
- For each processing step (POS, NER based on ne_chunk, custom NER, entity classification) list the main results (e.g. top entities)
- Compare results of entity classification approaches
  - nltk-based classification

    - wikipedia-based classification using nltk entities as the input
    - wikipedia-based classification using custom patterns as the input
- Provide the link to your implementation
    - You can use: https://gitlab.fit.cvut.cz [https://gitlab.fit.cvut.cz] or https://github.com/ [https://github.com/] or https://bitbucket.org/ [https://bitbucket.org/]
- Comment on
    - issues during the design/implementation
    - ideas for extensions/improvements/future work

/mnt/www/courses/MI-DDW.16/data/pages/hw/03/start.txt · Poslední úprava: 2017/03/24 07:54 autor: kuchajar