TP2 de SPL

Filipe and J.João

Contents

1 Emoji

- taxonomia
  - bandeiras, animais, emoção, objectos, paisagem
- features
  - emoção (feliz, raiva, preocupado,...) / polaridade
  - masc, feminino
- emojis compostos, apresentação preferenciais
- tokenisadores de textos com emojis
- normalização
- corpus (Ex: twitter da aula?)
- embeddings envolvendo emoji
- emoji textual " :) :-) :D "
- pares e conversores
  - emoji → português ":smile:"
  - emoji textual → emoji

2 Sentiment analysis

- baseada em regras
- listas
  - pal → polaridade (-101) ou prob {fea→ ...}
- texto → sent
- frases → sent
- lista de multiword elements ("deix.* muito a desejar","ponto forte")
- negadores (muito X, bem X, muito pouco X, não X, nada X, não tem X)
- Módulo parametrizado por algumas listas;
  - pal → polaridade, negadores, ...

  - sent: frase|texto → real
  - diagrama de sentimento: livro(=cap*) → (cap → sent)

3 Word embedings multilingua

- PT, EN
- transvec module (experimentar...)
- a partir de TMX (ver pasta TMX no git): Como?
  - tmx → frases embricadas → word-emb
- alguns testes a medir / validar os reultados?
  - similar(cão , dog)
  - casamentos(gato chair cat cão dog cadeira table mesa)
  - analogias?

4 TMX editor

- dada uma tmx (ver pasta TMX no git)
  - editor favorito
  - split dentro de unidades alinhadas # (dog, cat and bird = cão, gato e pássaro) # virar... (dog = cão), (cat = gato), (bird = pássaro)
  - corrigir alinhamentos interactivamente (Teclas programadas?, interfaces)

5 Melhorar LF-aligner

- baseado no Hun-aligner
- ferramentas auxiliares (html,pdf,docx → txt)
- recursos
- toolkit
- versão instalável! (pip install lf-aligner?)

6 Fluxo para preprocessamento e tratamento de corpus - multidocumento

- toolkit para preprocessar texto antes de
- spacy
  - (se pretendido) (V → Lema)
  - (se pretendido) NER
  - (se pretendido) remoção de pontuação, stopword, numeros
- aglutinação/marcação de termos multipalavra
- ... domain specific language para definição das reescritas de transformação (?)
- restauro de acentuações, capitalização (?)
- spell-checking (módulo spacy, hunspell?)
- outras...
- ... Wemb

7 Word embedings shell

gato : mamífero :: galinhas : ?

- importar algebra de gensim
  - similar
  - most similar
  - doesnt match
  - n_similar
- funções "nossas"
- caluladora

8 Multi-word terms processor (NL-flex)

```
DSL sistema de reescrita (padrão → acção python ou texto de substituição)

pequeno-almoço | café da manhã → refeiçao_PA
(\w+), (\w+) e (\w+) ==> insere_irmão( \1,\2,\3)
```