Preprocesamiento bioNLP

DEWY Team

09/10/18

1 Eliminar texto no informativo

- 1.1 Eliminar PMId. En shell:

```
cut -f3,4 sentences_RI_RIGC.txt
cut -f3,4 sentences_Other.txt
```

- 1.2. Remover referencias a tablas, figuras, artículos, revistas, etc. Se utilizó un script de python para ambos archivos:

```
import re
salida=open('/home/mescobar/Escritorio/Tercer_Semestre/Bioinfo
/senOth/Other_2.0.txt','w')

with open ('/home/mescobar/Escritorio/Tercer_Semestre/Bioinfo
/senOth/Other_0.5.txt','r') as archivo:

#Busca cualquier paréntesis con un número contenido, que
en la mayoría de los casos son referencias, "et al"
"Molecular microbiology" y lo sustituye por nada
    for line in archivo:
        line=re.sub(r"\(.*\d+.+?\)", "", line)
        line=re.sub(r"et al", "", line, flags=re.IGNORECASE)
        line=re.sub(r"molecular microbiology", "", line,
        flags=re.IGNORECASE)

        salida.write(line)
salida.close()
```

2 Unir en un mismo archivo

Unir ambos archivos libres de texto no informativo en el mismo archivo. En shell:

```
cat RI_2.0.txt Other2.0.txt > RI_Other_Cat.txt
```

3 Lematización y POS tagging

Lematización y POS tag del archivo con ambos tipos de oraciones. En CoreNLP:

```
./corenlp.sh -annotators tokenize,ssplit,pos,lemma,ner,parse,depparse

outputFormat conll -file  /home/mescobar/RI_Other_Cat.txt

-outputDirectory /home/mescobar/RI_Other_Cat.conll
```

Esto nos regreza la lematización y POS tagging para ambos tipos de oraciones, lo cual ya es informativo para poder vectorizarse.