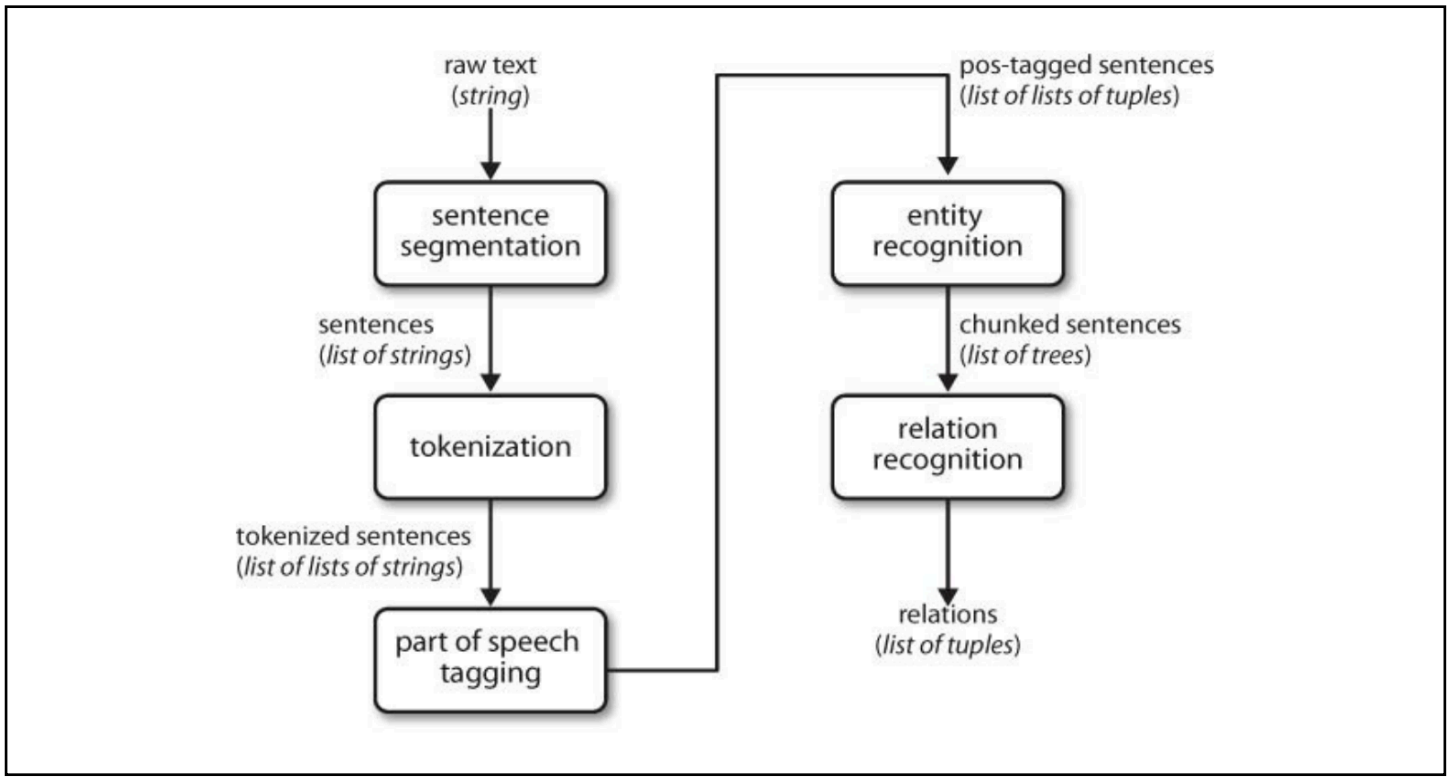
Student:	García González Aarón Antonio
Group:	3CV9
Learning unit:	Natural language processing
Teacher:	Kolesnikova Olga
Homework #12:	Information extraction
Date:	Monday, July 9th, 2020

We are going to use the next architecture for to do this activity:



To get sentences from text files, sentence segmentation, tokenization, I encoded the next function:

```
1.  def getSentences(PATH, OBJECT):
2.      result = []
3.
4.      for comment in os.listdir(PATH+OBJECT+"/"):
5.          path_file = PATH+OBJECT+"/"+comment
6.
7.          f = open(path_file, 'r', encoding = 'ISO-8859-1', errors="ignore")
8.          text = f.read()
9.          f.close()
10.
11.         text = re.sub('\n', ' ', text)
12.         text = text.strip()
13.
14.         sentences = nltk.sent_tokenize(text)
15.         sentences = [sentence.strip() for sentence in sentences if len(sentence) > 12]
16.         sentences = [nltk.word_tokenize(sentence) for sentence in sentences]
17.
18.         result += sentences
19.
20.     return result
```

To do POS tag, I used the “Stanford Log-linear Part-Of-Speech Tagger”, In this time the reason why I used this POS tagger instead of my POS tagger developed in class is only to know how functions this and so for next times can I use the best for my opinion or needs.

```
tagger = "/Users/aarongarcia/desktop/12_Info_extraction/stanford-tagger-4.0.0/models/spanish-ud.tagger"
jar = "/Users/aarongarcia/desktop/12_Info_extraction/stanford-tagger-4.0.0/stanford-postagger.jar"
reference = "https://nlp.stanford.edu/software/spanish-faq.shtml#tagset"
etiquetador = StanfordPOSTagger(tagger,jar)
saveFilePKL([etiquetador.tag(oracion) for oracion in oraciones],"etiquetas.pkl")
oraciones = getFilePKL("etiquetas.pkl")
```

This tool has only one issue, is so slow, that is the reason why I use pickle files.
The first point of this homework is getting noun phrases, I obtained a text file, I show some of these phrases:

```
(NP hoteles/NOUN)
(NP estrellas/NOUN)
(NP mejores/ADJ)
(NP no/ADV creo/VERB)
(NP este/DET hotel/NOUN)
(NP Las/DET habitaciones/NOUN)
(NP pequeñas/ADJ)
(NP no/ADV tienen/VERB)
(NP camas/NOUN de/ADP matrimonio/NOUN)
(NP terraza/NOUN)
(NP decoradas/ADJ)
(NP manera/NOUN antigua/ADJ)
(NP algunas/DET grietas/NOUN)
(NP Los/DET mandos/NOUN)
(NP la/DET tele/NOUN)
(NP no/ADV funcionan/VERB)
(NP no/ADV tienen/VERB)
(NP canales/NOUN)
(NP no/ADV comentar/VERB)
(NP además/ADV estan/VERB)
(NP sucias/NOUN)
(NP los/DET baños/NOUN)
(NP la/DET habitación/NOUN)
(NP Las/DET instalaciones/NOUN)
(NP viejas/ADJ y/CCONJ muy/ADV descuidadas/ADJ)
(NP el/DET gimnasio/NOUN)
(NP la/DET piscina/NOUN climatizada/ADJ solo/ADJ)
(NP la/DET tarde/NOUN)
(NP masajes/NOUN)
(NP no/ADV habia/VERB)
(NP masajista/NOUN)
(NP mala/ADJ)
(NP el/DET entorno/NOUN)
(NP bonito/ADJ)
(NP personal/NOUN del/ADP hotel/NOUN)
(NP amable/ADJ)
(NP las/DET zonas/NOUN comunes/ADJ)
(NP desayuno/NOUN tipo/NOUN buffet/NOUN)
(NP lo/DET mejor/ADJ)
(NP una/DET revisión/NOUN)
(NP calidad/NOUN del/ADP hotel/NOUN)
(NP las/DET estrellas/NOUN)
(NP caro/ADJ)
(NP euros/noche/NOUN)
(NP Realmente/ADV conozco/VERB)
(NP hoteles/NOUN)
(NP estrellas/NOUN)
(NP una/DET pasada/ADJ)
(NP Queria/PROPN)
(NP mi/DET novio/NOUN)
(NP un/DET finde/NOUN)
(NP un/DET hotel/NOUN)
(NP estrellas/NOUN)
(NP una/DET gran/ADJ desilusión/NOUN)
(NP Pocos/DET establecimientos/NOUN hoteleros/ADJ)
(NP Madrid/PROPN)
(NP mejor/ADJ situación/NOUN)
(NP Tryp/PROPN)
(NP Ambassador/PROPN)
(NP establecimiento/NOUN)
(NP estrellas/NOUN)
(NP la/DET cadena/NOUN)
(NP Sol-Meliá/PROPN)
(NP ubicado/ADJ)
(NP Cuesta/PROPN)
(NP Santo/PROPN)
(NP Dominfgo/PROPN)
(NP un/DET paso/NOUN del/ADP Palacio/PROPN)
(NP Real/PROPN)
(NP Senado/PROPN)
(NP Ópera/PROPN)
(NP la/DET Gran/ADJ)
(NP Vía/PROPN)
(NP Plaza/PROPN)
(NP España/PROPN)
(NP el/DET antiguo/ADJ)
(NP Palacio/PROPN)
(NP Duques/PROPN)
(NP Granada/PROPN)
(NP Ega/PROPN)
```

The grammar that I used to do this:

```
grammar_NP = r""NP:
                {<DET><NOUN><ADP><PROPN>}
                {<NOUN><NOUN><NOUN>*}
                {<NOUN><ADP><NOUN>}
                {<DET>?<ADJ><NOUN>}
                {<DET>?<NOUN><ADJ>*}
                {<DET><ADJ>}
                {<ADV><VERB><NOUN>?}
                {<DET><ADV>?<ADJ>}
                {<ADJ><CCONJ><ADV>?<ADJ>}
                {<ADJ>?<NOUN>?}
                {<PROPN><ADJ>*}
                ""
```

The second point is getting sentence extraction, the grammar that I used to do this:

```
grammar_CHUNK = r""CHUNK:
                {<DET><NOUN><AUX><ADJ>}
                {<NOUN><ADP><NOUN>}
                {<ADJ><ADP><NOUN><ADJ>}
                {<DET>?<NOUN><VERB><ADJ>}
                {<DET><NOUN><ADJ>}
                {<ADV><VERB><NOUN>}
                {<DET><NOUN><ADV><VERB><NOUN|VERB>*}
                {<NOUN>+<ADP|DET>+<ADJ|NOUN>+}
                {<DET>?<NOUN><ADP><DET>?<NOUN|VERB|ADJ>+}
                {<DET><NOUN><ADP><DET><NOUN>}
                {<DET><NOUN><AUX><DET>?<NOUN>}
                {<PROPN><ADP><PROPN|NOUN>+}
                {<DET><NOUN><VERB><DET>+<NOUN>}
                ""
```

I obtained a text file; I show some of these phrases:

```
(CHUNK Las/DET habitaciones/NOUN son/AUX pequeñas/ADJ)
(CHUNK camas/NOUN de/ADP matrimonio/NOUN)
(CHUNK decoradas/ADJ de/ADP manera/NOUN antigua/ADJ)
(CHUNK la/DET tele/NOUN no/ADV funcionan/VERB)
(CHUNK además/ADV estan/VERB sucias/NOUN)
(CHUNK Las/DET instalaciones/NOUN estan/VERB viejas/ADJ)
(CHUNK la/DET piscina/NOUN climatizada/ADJ)
(CHUNK no/ADV habia/VERB masajista/NOUN)
(CHUNK el/DET entorno/NOUN es/AUX bonito/ADJ)
(CHUNK personal/NOUN del/ADP hotel/NOUN)
(CHUNK las/DET zonas/NOUN comunes/ADJ)
(CHUNK desayuno/NOUN tipo/NOUN buffet/NOUN de/ADP lo/DET mejor/ADJ)
(CHUNK calidad/NOUN del/ADP hotel/NOUN)
(CHUNK Realmente/ADV conozco/VERB hoteles/NOUN)
(CHUNK novio/NOUN un/DET finde/NOUN)
(CHUNK Pocos/DET establecimientos/NOUN hoteleros/ADJ)
(CHUNK estrellas/NOUN de/ADP la/DET cadena/NOUN)
(CHUNK Cuesta/PROPN de/ADP Santo/PROPN Dominfgo/PROPN)
(CHUNK Plaza/PROPN de/ADP España/PROPN)
(CHUNK Duques/PROPN de/ADP Granada/PROPN)
(CHUNK su/DET estructura/NOUN original/ADJ)
(CHUNK un/DET establecimiento/NOUN muy/ADV frecuentado/VERB)
(CHUNK personajes/NOUN de/ADP la/DET pequeña/ADJ pantalla/NOUN)
(CHUNK habitaciones/NOUN para/ADP sus/DET estancias/NOUN)
(CHUNK zona/NOUN de/ADP teatros/NOUN)
(CHUNK Tv/NOUN de/ADP pago/NOUN)
```