Introducción al Procesamiento de Lenguaje Natural

Grupo PLN – InCo

Tokenización
&
Normalización
&
Segmentación

Unidades de texto

¿Cuáles son las unidades independientes más pequeñas del texto?

-Segmento del discurso unificado habitualmente por el acento, el significado y pausas potenciales inicial y final

- Aquellas que puede existir en forma libre y que conforma el enunciado o mensaje lingüístico. Están dotadas de significado léxico o gramatical, según el caso

*Son entonces **signos lingüísticos***

Palabras

- La **palabra** es un conjunto o secuencia de sonidos articulados que se pueden representar gráficamente con letras, y por lo general, asocian un significado

- Mínima unidad con significado (Aristóteles)

Unidad de texto = Palabra

¿Qué es una palabra?

José Arcadio Buendía conversó con Prudencio Aguilar hasta el amanecer. Pocas horas después, estragado por la vigilia, entró al taller de Aureliano y le preguntó: '¿Qué día es hoy?' Aureliano le contestó que era martes. 'Eso mismo pensaba yo', dijo José Arcadio Buendía. 'Pero de pronto me he dado cuenta de que sigue siendo lunes, como ayer. Mira el cielo, mira las paredes, mira las bego-nias. También hoy es lunes.' Acostumbrado a sus manías, Aureliano no le hizo caso. Al día siguiente, miércoles, José Arcadio Buendía volvió al taller. 'Esto es un desastre –dijo–. Mira el aire, oye el zumbido del sol, igual que ayer y antier. También hoy es lunes.' Esa noche, Pietro Crespi lo encontró en el corredor, llorando con el llantito sin gracia de los viejos, llorando por Prudencio Aguilar, por Melquíades. *(Cien años de soledad – Gabriel García Márquez)*

Palabras, Tipos y Tokens

- Palabras: unidades que hay en un corpus o en un vocabulario

- Tipos (word types): unidades *distintas* que hay en un corpus o en un vocabulario

- Tokens: son las instancias de los tipos en un corpus o en un vocabulario

Corpus

¿Qué es un Corpus?

- Es una colección de material lingüístico de ejemplos reales de uso de la lengua

- Es de utilidad en diferentes áreas, principalmente en lingüística computacional y lingüística teórica

Corpus

¿Cómo se construye un corpus?

- Recopilación de un conjunto de documentos
- Hay que definir las características deseadas:
  - escrito / oral
  - idioma (un idioma o multilingüe)
  - tipo de texto (prensa, literario, científico, ...)
  - dominio (arte, lingüística, deportes, informática, ...)
  - anotado / no anotado (conjunto de etiquetas)
  - ...

Corpus

Algunos corpus en inglés

- Brown Corpus
- Penn Treebank
- PropBank

Algunos corpus en español

- CREA y CORDE (RAE)
- Corpus del español de Mark Davies
- Ancora
- Adesse
- “Nuestros”

Corpus, palabras y tokens

- Corpus CREA
  (Corpus de Referencia del Español Actual) desde el año 1975 al 2004

  - *140.000 documentos*
  - *737.799 palabras*
  - *125 millones de tokens*

Corpus, palabras y tokens

La bella y graciosa moza, marchose a lavar la ropa

¿Cuántas palabras? ¿Cuántos tipos? ¿Cuántos tokens?

Corpus, palabras y tokens

La bella y graciosa moza, marchose a lavar la ropa

¿Cuántas palabras?	¿Cuántos tipos?	¿Cuántos tokens?
10	9	10

Corpus, palabras y tokens

… entró al taller de Aureliano y le preguntó: '¿Qué día es hoy?' Aureliano le contestó que era martes. 'Eso mismo pensaba yo', dijo José Arcadio Buendía. 'Pero de pronto me he dado cuenta de que sigue siendo lunes, como ayer. Mira el cielo, mira las paredes, mira las begonias. También hoy es lunes.' Acostumbrado a sus manías, Aureliano no le hizo caso. Al día siguiente, miércoles, José Arcadio Buendía volvió al taller. 'Esto es un desastre –dijo–. Mira el aire, oye el zumbido del sol, igual que ayer y …

¿Cuántas palabras? ¿Cuántos tipos? ¿Cuántos tokens?

Corpus, palabras y tokens

… entró al taller de Aureliano y le preguntó: '¿Qué día es hoy?' Aureliano le contestó que era martes. 'Eso mismo pensaba yo', dijo José Arcadio Buendía. 'Pero de pronto me he dado cuenta de que sigue siendo lunes, como ayer. Mira el cielo, mira las paredes, mira las begonias. También hoy es lunes.' Acostumbrado a sus manías, Aureliano no le hizo caso. Al día siguiente, miércoles, José Arcadio Buendía volvió al taller. 'Esto es un desastre –dijo–. Mira el aire, oye el zumbido del sol, igual que ayer y …

¿Cuántas palabras? ¿Cuántos tipos? ¿Cuántos tokens?

Tokenización

- Identificar las palabras (*tokens*) en un texto
- Los espacios separan tokens, pero…
  - escribió “El príncipe feliz”
  - el 24 de agosto de 1889
  - Hay 10,000 razones para no creer
  - Lo busqué en http://rulzindeed.blogspot.com
  - Je t’aime, rock’n’roll
  - New York
  - estado del arte
  - al, del
- El estándar de tokenización del Penn Treebank

Tokenización

- El chino y el japonés no marcan los límites de palabras

  小時候沒有人選擇我踢足球

  De chico nadie me elegía para jugar al fútbol

- En alemán algunas palabras compuestas pueden escribirse todas juntas
  - hora pico → hauptverkehrszeit
  - a veces → manchmal
  - jugo de frutas → fruchtsaft

Tokenización

Algoritmo MaxMatch

- Basado en una lista de palabras
- Comienza al principio de la entrada
- Elige siempre la palabra más larga en la posición actual de la entrada
- Si no encuentra ninguna palabra, se crea una palabra de una letra
- Avanza al puntero a la primera posición luego de la palabra encontrada

MaxMatch

Entrada:

"mesacadelacanchasinmotivo"

Avance (en negrita la entrada ya analizada)

mesacadelacanchasinmotivo

mesacadelacanchasinmotivo

mesacadelacanchasinmotivo

mesacadelacanchasinmotivo

...

Salida:

[Mesa, ca, de, la, cancha, sin, motivo] ó

[Mesa, ca, del, a, cancha, sin, motivo] ó

[Mes, aca, de, la, cancha, sin, motivo]

MaxMatch

Algoritmo

```
function MAXMATCH(sentence, dictionary D) returns word sequence W

    if sentence is empty
        return empty list
    for i ← length(sentence) downto 1
        firstword = first i chars of sentence
        remainder = rest of sentence
        if InDictionary(firstword, D)
            return list(firstword, MaxMatch(remainder,dictionary) )

    # no word was found, so make a one-character word
    firstword = first char of sentence
    remainder = rest of sentence
    return list(firstword, MaxMatch(remainder,dictionary D) )
```

Evaluación

¿Cómo evaluamos un tokenizador?

- Input 1: Nuestra segmentación
  [Mesa, ca, de, la, cancha, sin, motivo]
- Input 2: La segmentación correcta (gold standard)
  [Me, saca, de, la, cancha, sin motivo]

Word Error Rate (Distancia Mínima de Edición)

¿cuántas palabras deben insertarse, borrarse o sustituirse para ir de Input 1 a Input 2?

(En el ejemplo anterior, 2)

Evaluación

¿Cómo evaluamos un tokenizador?

- Para lenguajes como el español y el inglés no funciona muy bien….
- pero si funciona bastante bien para el chino
- Hoy en día algoritmos de tokenización probabilistas funcionan mejor…

Normalización

Normalización: llevar las palabras a un formato estándar

- Llevar los números a un formato único
- URLs y otras formas con estructura
- Detección de entidades con nombre
- Llevar todo a minúsculas/mayúsculas
- ...

Tradicionalmente, para la tokenización y normalización se han utilizado técnicas modeladas con autómatas finitos.

Segmentación de Oraciones

Punkt Sentence Tokenizer:

- Toma un corpus y lo divide en oraciones
- Problemas con el “.”
  - Identifica candidatos a abreviaturas
    + Siempre terminan en punto
    + En general, son cortas
    + En general, tienen puntos internos
  - Fin de oración
  - Puntos interiores a números
- Utiliza heurísticas a nivel de token para corregir:
  + Mayúsculas
  + Colocaciones
  + "Iniciales frecuentes"

Segmentación de Oraciones

Punkt Sentence Tokenizer

- Algoritmo no supervisado y entrenado en una gran colección de textos antes de ser utilizado

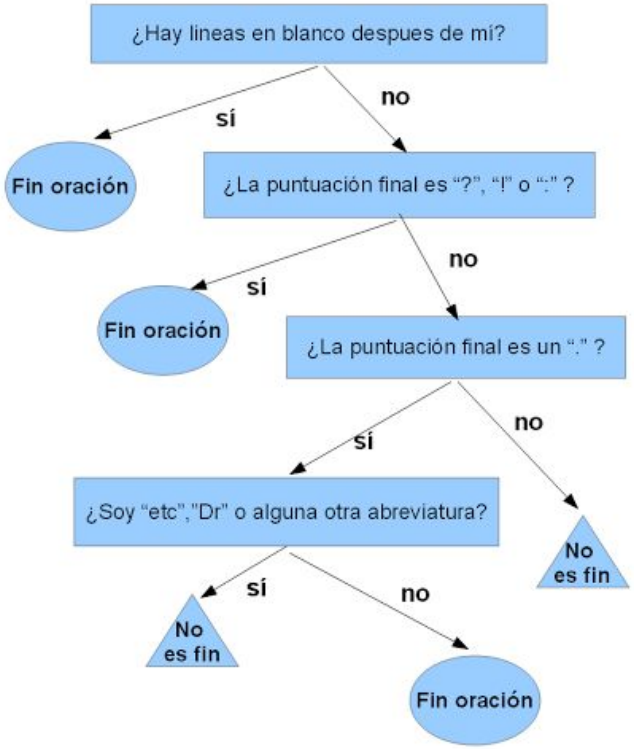
- Cuenta la frecuencia de cada palabra

- Criterio: si el signo de puntuación no es parte de una abreviatura, entonces separa oraciones

- En *nltk* está el paquete pre-entrenado para el inglés (nltk.tokenize.*punkt*)

Segmentación de Oraciones

Árboles de decisión



Referencias

- J.Martin & D.Jurafsky. Speech and Language Processing – Chapter 2 (2014)
- Greffenstette & Tapainanen – What is a word, what is a sentence? Problems of tokenization (1994)
- Kiss & Strunk – Unsupervised Multilingual Sentence Boundary Detection (2006)
- Steven Bird – Natural Language Processing with Python – Chapter 3: "Processing Raw Text" (2009)