Introducción a Natural Language Processing

Procesamiento Básico de Texto

Ronald Cárdenas Acosta

Setiembre, 2016

Outline

1. Tokenization
2. Normalización de Palabras
3. Expresiones Regulares
4. NLTK: Natural Language Toolkit

Outline

1. Tokenization
2. Normalización de Palabras
3. Expresiones Regulares
4. NLTK: Natural Language Toolkit

Tokenización o segmentación

- Se debe decidir los casos de separación de palabras u oraciones.
- Puede ser ambigüo:
  Distinguir entre punto de abreviación (*Mr.*) y punto de final de oración (*... temina oración.*)
- Requiere de modelos estadísticos para lenguajes basados en idiogramas (chino, japonés, coreano), en los que no se utiliza espacio entre palabras.

Outline

1 Tokenization

2 Normalización de Palabras

3 Expresiones Regulares

4 NLTK: Natural Language Toolkit

Normalización de palabras (1)

- Consiste en eliminar las particulas de inflexión (morfemas de conjugación):
- Se puede definir casos de equivalencia de términos (*S.U.N.A.T* y *SUNAT*)
- Stemming: reemplaza la palabra por su raíz.
- Lemmatización: reemplaza la palabra por su lexema o lemma (requiere el uso de lexicon)

Normalización de palabras (2)

- Minimizado de palabras: Los usuarios tienden a usar minúsculas. Excepciones deben ser obviadas (nombres propios, abreviaciones)
- Corrección Ortográfica (*spell checker*): se basa en algoritmo Edit Distance
- Eliminación de ruido:
  Consiste en reemplazar palabras menos frecuentes por la etiqueta OTROS

Lidiando con palabras de baja frecuencia

Reemplazar palabra de baja frecuencia con etiqueta según forma ortografíca:

- allCaps: todo mayúsculas
- initCap: primera letra mayúscula
- Number: si es numero
- Phono: número telefónico
- Date/Hour: fecha u hora
- URL: link de algún website

Lidiando con palabras de baja frecuencia

Figure: Etiquetas para palabras de baja frecuencia, preprocesado previo a NER

[Bikel et. al 1999] **(named-entity recognition)**

Word class	Example	Intuition
twoDigitNum	90	Two digit year
fourDigitNum	1990	Four digit year
containsDigitAndAlpha	A8956-67	Product code
containsDigitAndDash	09-96	Date
containsDigitAndSlash	11/9/89	Date
containsDigitAndComma	23,000.00	Monetary amount
containsDigitAndPeriod	1.00	Monetary amount, percentage
othernum	456789	Other number
allCaps	BBN	Organization
capPeriod	M.	Person name initial
firstWord	first word of sentence	no useful capitalization information
initCap	Sally	Capitalized word
lowercase	can	Uncapitalized word
other	,	Punctuation marks, all other words

Outline

1 Tokenization

2 Normalización de Palabras

3 Expresiones Regulares

4 NLTK: Natural Language Toolkit

Expresiones Regulares

- Meta-lenguaje para buscar patrones de texto en base a reglas hechas a mano.
- Usadas extensamente en la industria
- Ejemplos de aplicación:
  - Extraer números telefónicos y fechas
  - Limpiar texto web (saltos de linea, espacios extra, html tags)
- Libreria en Python: **re**
  - `https://docs.python.org/2/howto/regex.html`
  - `https://docs.python.org/2/library/re.html`

Outline

1 Tokenization

2 Normalización de Palabras

3 Expresiones Regulares

4 NLTK: Natural Language Toolkit

NLTK: Natural Language Toolkit

- Librería más completa para procesar texto en Python.
- La mayoría de funciones y modelos son independientes del lenguaje a procesar.
- Contiene extractos de corpus standard para las aplicaciones más populares de NLP (tagging, parsing, machine translation, entre otros).