Lingüística Computacional

Tokenizing

Antonio Frías Delgado

Área de Lógica y Filosofía de la Ciencia
Departamento de Historia, Geografía y Filosofía
Universidad de Cádiz

Segmentación: Bibliografía

Segmentación: Bibliografía

- What is a word, What is a sentence? Problems of Tokenization
  Gregory Grefenstette, Pasi Tapanainen

Segmentación: Bibliografía

- What is a word, What is a sentence? Problems of Tokenization
  Gregory Grefenstette, Pasi Tapanainen
- Text preprocessing
  David Palmer

Segmentación: Bibliografía

- What is a word, What is a sentence? Problems of Tokenization
  Gregory Grefenstette, Pasi Tapanainen
- Text preprocessing
  David Palmer
- Tokenisation and sentence segmentation
  David Palmer

Segmentación: Recursos

Segmentación: Recursos

NLTK

Segmentación: Recursos

NLTK

- `import nltk`

Segmentación: Recursos

NLTK

- `import nltk`
- `nltk.word_tokenize(texto,language='spanish')`

Segmentación: Recursos

NLTK

- `import nltk`
- `nltk.word_tokenize(texto,language='spanish')`
- `nltk.wordpunct_tokenize(texto)`

Segmentación: Recursos

NLTK

- `import nltk`
- `nltk.word_tokenize(texto,language='spanish')`
- `nltk.wordpunct_tokenize(texto)`
- `nltk.sent_tokenize(texto,language='spanish')`

NLTK: Brown Corpus

- `import nltk`

NLTK: Brown Corpus

- `import nltk`
- `from nltk.corpus import brown`

NLTK: Brown Corpus

- `import nltk`
- `from nltk.corpus import brown`
- `text_brown=brown.raw()`

NLTK: Brown Corpus

- `import nltk`
- `from nltk.corpus import brown`
- `text_brown=brown.raw()`
- `oraciones_brown=brown.sents()`

NLTK: Brown Corpus

- `import nltk`
- `from nltk.corpus import brown`
- `text_brown=brown.raw()`
- `oraciones_brown=brown.sents()`
- `palabras_brown=brown.words()`