Procesamiento del Lenguaje Natural – PLN

El Procesamiento del Lenguaje Natural es una rama de conocimiento de la Inteligencia Artificial que pretende conseguir que una máquina comprenda lo que expresa una persona mediante el uso de un lenguaje natural (inglés, español, chino...). Los lenguajes naturales pueden expresarse de forma oral (mediante la voz), o por escrito (un texto). El Procesamiento del Lenguaje Natural está mucho más avanzado en el tratamiento de textos escritos. Las posibilidades de los modelos a aplicar se enfocan no solo a la comprensión del lenguaje de por sí, sino a aspectos relacionados con la organización de la información, así como a la importancia de los conceptos.

Para transformar un texto escrito en un lenguaje natural a algo interpretable por un ordenador se pueden usar diferentes tipos de análisis, pero de forma básica intentaremos una aproximación en la que al menos un programa sea capaz de etiquetar las palabras y producir estadísticas de frecuencias en un primer paso. Esto suele ser muy útil, por ejemplo, en disciplinas relacionadas con la medicina clínica donde el profesional tradicionalmente toma notas manuscritas o en ordenador de la situación clínica del paciente y luego esta información debería ser útil para el diagnóstico (en la historia del paciente).

Nubes de etiquetas

Igualmente, en otras áreas se usa en esta primera fase, para producir diagramas/mapas de resumen de conceptos con diferentes propósitos. En este caso, mapas/nubes de palabras o etiquetas, como representación visual de las palabras que conforman un texto, en donde el tamaño es mayor para las palabras que aparecen con más frecuencia.

Uno de sus usos principales es la visualización de las etiquetas de un sitio web, de modo que los temas más frecuentes en el sitio se muestren con mayor prominencia. También en medicina o cualquier otra disciplina.

Las etiquetas son palabras clave que suelen estar ordenadas alfabéticamente o, en ocasiones, agrupadas semánticamente. La importancia de una etiqueta se muestra con el tamaño de la fuente y/o color.

Enunciado: Diseñar un programa para dado un texto almacenado en un fichero.txt, construir un mapa de etiquetas de palabras.

Implementación

`Texto1.m`: script que realiza el PLN y genera la nube de etiquetas

`FI_1_1.txt`: fichero de texto que contiene la primera pregunta del tema 1 de Fundamentos de Informática.

Referencias

Mariani, Joseph; Francopoulo, Gil; Paroubek, Patrick; Vernier, Frédéric (2019), «The NLP4NLP Corpus (I): 50 Years of Research in Speech and Language Processing», Frontiers in Research Metrics and Analytics

Hassan-Montero, Y., Herrero-Solana, V. Improving Tag-Clouds as Visual Information Retrieval Interfaces. InSciT 2006: Mérida, Spain. October 25–28, 2006.

Hassan-Montero, Y., Herrero-Solana, V., Guerrero-Bote, V.; Usabilidad de los tag-clouds: estudio mediante eye-tracking. SCIRE, Vol. 16, n. 1, 2010, pp. 15-33.

NUBE de PALABRAS según su frecuencia

```
close all;
clear all;
```

Lectura de un fichero de texto a analizazr

```
texto = fileread('FI_1_1.txt');
disp('El comienzo del texto del fichero es: ');
texto(1:300)
```

```
El comienzo del texto del fichero es:

ans =

    'La Informática y la Ingeniería
     Es necesario comenzar intentando dar una explicación al nacimiento de esta nueva
     ciencia denominada Informática. La situación es fácil de imaginar: vivimos en un
     mundo en el que estamos continuamente bombardeados por todo tipo de
     informaciones, que en muchos de los'
```

Se convierte el texto a string

Se separa en líneas

```
texto = string(texto);
texto = splitlines(texto);
disp('El comienzo es : ');
texto(1:5)
```

```
El comienzo es :

ans =

  5×1 string array

    "La Informática y la Ingeniería"
    "Es necesario comenzar intentando dar una explicación al nacimiento de esta nueva"
    "ciencia denominada Informática. La situación es fácil de imaginar: vivimos en un"
    "mundo en el que estamos continuamente bombardeados por todo tipo de"
    "informaciones, que en muchos de los casos varían con el tiempo. Vivimos por tanto,"
```

Se sustituyen los signos de puntación por espacios

```
p = ["." "?" "!" "," ";" ":"];
texto = replace(texto,p," ");
texto(1:5)
```

```
ans =

  5×1 string array

    "La Informática y la Ingeniería"
```

```
    "Es necesario comenzar intentando dar una explicación al nacimiento de esta nueva"
    "ciencia denominada Informática La situación es fácil de imaginar vivimos en un"
    "mundo en el que estamos continuamente bombardeados por todo tipo de"
    "informaciones  que en muchos de los casos varían con el tiempo  Vivimos por tanto "
```

Se divide el texto en un string array con palabras individuales

Se unen todas y se divide según los espacios que se encuentren

```
texto = join(texto);
texto = split(texto);
texto(1:15)
```

```
ans =

  15×1 string array

    "La"
    "Informática"
    "y"
    "la"
    "Ingeniería"
    "Es"
    "necesario"
    "comenzar"
    "intentando"
    "dar"
    "una"
    "explicación"
    "al"
    "nacimiento"
    "de"
```

Borrar palabras de longitud menor que 5

```
texto(strlength(texto)<5) = [];
texto(1:15)
```

```
ans =

  15×1 string array

    "Informática"
    "Ingeniería"
    "necesario"
    "comenzar"
    "intentando"
    "explicación"
    "nacimiento"
    "nueva"
    "ciencia"
    "denominada"
    "Informática"
```

```
"situación"
"fácil"
"imaginar"
"vivimos"
```

Se convierte el texto un array de categorías

Se representa según su frecuencia

```
C = categorical(texto);
figure, wordcloud(C);
title("Mapa de frecuencia de palabras");
```

Published with MATLAB® R2018b