Propagación de sentimiento con técnicas semi-supervisadas

Karen Haag

FaMAF, Universidad Nacional de Córdoba

kah0112@famaf.unc.edu.ar

22 de Noviembre, 2017

Overview

Aprendizaje semi-supervisado

El aprendizaje semi-supervisado es una clase de técnicas de aprendizaje automático que utiliza datos de entrenamiento tanto etiquetados como no etiquetados.

Es muy útil cuando tenemos una pequeña cantidad de datos etiquetados junto a una gran cantidad de datos no etiquetados.

Descripción de la tarea

- Obtener un lexicón anotado con sentimiento
- Obtener dos corpus de tweets: Uno no anotado, y otro anotado.
  - Pre-procesar los corpus
- Utilizaremos 3 técnicas semi-supervisadas para agregar más palabras al corpus:
  1. Propagación de sentimiento con grafos.
  2. Selección por fuerte información mútua.
  3. Selección por votación.
- Clasificación de tweets por clase mayoritaria
- Evaluación de las técnicas semi-supervisadas

Lexicón anotado

Se utilizó un lexicón anotado continuamente (-1, 1) con sentimiento de 11383 palabras. Lo pueden descargar de: ML-SentiCon

Problemas: Algunas palabras aparecen como "Positivas" y "Negativas" a la vez.
Por ejemplo: "astuto","salvador","cerdo", "ebrio"

Solución: Eliminarlas de la clase con menos probabilidad de ocurrir.

Corpus de twitter anotado y no anotado

Se utilizó un corpus de twitter en español no anotado de 60798 tweets para aplicar las diferentes técnicas.
También uno anotado para testing y para definición de polaridad en palabras ambigüamente asociadas.

Ambos corpus los pueden obtener de : SEPLN-TASS15

Pre-procesamiento de los datos: Se tokenizó en primer instancia para observar los datos

Limpieza del corpus: Se eliminaron hashtag, mensiones, emoticonos y enlaces para tener datos mas representativos en el corpus

Propagación de sentimiento con grafos

Se utiliza la librería Label Propagation de Scikit Learn.

Esta librería toma como parámetro una matriz [n-samples, m-features]y otra con [n-labels]y retorna una matrix [n-samples, n-samples]con los pesos que se han propagado entre todos los nodos.

Para distinguir entre ejemplos etiquetados y no etiquetados se le asigna a los no etiquetados el valor "-1" en su lugar correspondiente de la matriz de labels.

Para más información sobre Label Propagation puede ver: Label propagation semisupervised learning with applications to nlp

Selección por fuerte información mútua

Utilizamos el corpus no anotado de twitter para adquirir nuevas palabras que le podamos asociar sentimiento.

1. Analizamos las palabras que se encuentren en un contexto k de las palabras semilla.
   - Si un tweet no tiene ninguna palabra semilla en su contenido, lo descartamos como ejemplo de aprendizaje en esta iteración.
2. Agregamos al corpus las palabras fuertemente asociadas a palabras con algún sentimiento
3. Iterar a 1.

Selección por votación

1. Para cada tweet, le asigno clase positiva o negativa por votación
   - Si un tweet no tiene ninguna palabra semilla en su contenido, lo descartamos como ejemplo de aprendizaje en esta iteración.
2. Incorporo las palabras que han sido asignadas a oraciones positivas con valor 1 y las palabras que han sido asignadas como negativas con -1.
   - Si una palabra es asociada a ambas clases se le asigna un valor continuo entre 1 y -1 que es una función de la proporción de veces que ocurrió en oraciones positivas o negativas
3. Iteramos a 1

Clasificador por votación

Construimos un clasificador que trabaja por votación para la evaluación de las técnicas: Dado un tweet le asigna una clase "pesando" la polaridad de sus palabras.

Análisis de las técnicas

Utilizamos un corpus anotado con sentimiento de twitter para evaluar las tres técnicas utilizando el clasificador.

Plan de experimentos

Por cada técnica:

- Se utilizará el clasificador con el lexicón resultante final para clasificar tweets.
- Se comparará predicción con el tag de los tweets
- Finalmente se comparará efectividad entre las diferentes técnicas semi-supervisadas.

Para la técnica de propagación por grafo se recomienda implementar diferentes cantidad de iteraciones para ver cual es la más efectiva.

Mejoras futuras

- Utilizar nuevas técnicas semi-supervisadas.
- Propagar sentimientos a hashtags
- Eliminar Stopwords
- Utilizar corpus diferentes a twitter para entrenar

Preguntas?

Gracias!