**Experimentos sobre corpus de opiniones**

**Búsqueda de contextualización**

**Objetivo**: Desarrollar un sistema detección de sentimientos que trabaje contextualizando las opiniones en foros, con diferentes tópicos, usando la bolsa de palabras de Bing Liu y explorando otras opciones en el análisis como Naive Bayes en la clasificación y en el pre proceso mejora de los datos aplicando stop words y otras heurísticas propias que se acomodan a el dominio de la información.

**Resumen**

Se tiene una serie de archivos que contienen, puntuaciones sobre la descripción del producto en la tienda de Amazon sobre los siguientes productos:

1. digital camera: Canon G3

2. digital camera: Nikon coolpix 4300

3. cellular phone: Nokia 6610

4. mp3 player: Creative Labs Nomad Jukebox Zen Xtra 40GB

5. dvd player: Apex AD2600 Progressive-scan DVD player

Cuenta con notación dentro de las características para identificar:

Donde en el pre proceso se descarta otros identificadores y se centran en indicadores de valoración positivos[+n] y negativos[-n]

En el pre proceso se entregan tres elementos, (1) nombre de archivo, (2) características con puntuación las que existan y (3) frase de opinión.

Por ejemplo:

corpus2/Nokia 6600.txt.G,audio quality[+3], 1) Great audio quality overall.

corpus2/Nokia 6600.txt.G,audio quality[+3],Handset quality is great as is the speaker phone quality and the headset audio quality.

Se aplica el preproceso que limpia el corpus, remplaza de los conjuntos de características cuando se tienen +/+=+ , -/-=- y con +/-=- , donde al + se remplaza como positive y el – como negative. Finalmente queda asi :

[('im a more happier person after discovering the i/p button ! ', 'positive'),

("but if you 're looking for my opinion of the apex dvd player i love it ! ", 'positive'),

('it practically plays almost everything you give it ', 'positive'),

("i 've had the player for about 2 years now and it still performs nicely with the exception of an occasional wwhhhrrr sound from the motor ", 'negative'),

('for the price it is a well spent investment ! ', 'positive')]

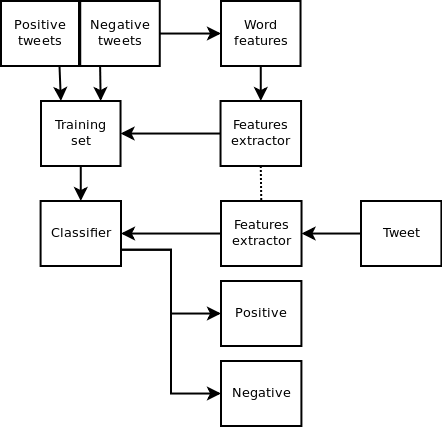
Desarrollo

Se desarrollaron varias aproximaciones,

Una de las aproximaciones,

1. Tokenizar las frases de las opiniones
2. Separar el corpus en porcentajes a training y clasificación(test)
3. Extracción de características basados en la frecuencia de aparición de palabras TF
4. Clasificador basado en Naive Bayes, sabiendo que tenemos un 50% de probabilidad de ser 0.5 positivo o un 0.5 negativo.

Explicación aproximación 1



Aproximación 2

Tokenizar las opiniones

Pasar stop words / Pasar stop words lista preprocesada POS

Separar el corpus en porcentajes a training y clasificación

Extracción de características basados en la frecuencia de aparición TF

Clasificador basado en Naive Bayes

Explicación aproximación 2

POS tag JJR //adjective, comparative

POS tag RBR //adverb, comparative

POS tag JJS //adjective, superlative

POS tag RBS //adjective, superlative

Aproximación 3 Clasificador Basado En SVM

Tokenizar las opiniones

Pasar stop words / Pasar stop words lista preprocesada POS

Separar el corpus en porcentajes a training y clasificación

Extracción de características basados en la frecuencia de aparición TF

Clasificador basado en SVM sabiendo que tenemos un 50% de probabilidad de ser 0.5 positivo o un 0.5 negativo.

Explicación aproximación 3

POS tag JJR //adjective, comparative

POS tag RBR //adverb, comparative

POS tag JJS //adjective, superlative

POS tag RBS //adjective, superlative

**Resultados**

**Conclusiones**.

Algunos de los problemas que se presentan es el texto escrito no es comprensible por que las personas que escriben usan abreviaturas o simplemente esta escrito mal. Probablemente en esta situación se podría pasar un filtro corrector y desambiguar algunas palabras sobre el contexto.

Formular polaridad automáticamente, “in case have a positive % and negative % with the same context determine polarity in the context” well if you dont use sentiwordnet

>Tomar un foro de opinión donde exista una calificación global y cada opinión no este clafificada mirando si existe una correlacion entre el termino de frecuencia de palabras escritas en las opiniones, entrenando desde relación a sentiwordnet y POS

otra idea es tomar las opiniones de una noticia de un suceso de frecuencia y mirar como a sido con el tiempo sobre el numero de opiniones positivas o negativas, por ejemplo comentarios sobre noticias económicas, deportivas, políticas.

El método de bag of words permite que el sistema funcione independiente del lenguaje con el que funciones el clasificador.

**Referencias**

Minqing Hu and Bing Liu. "Mining and summarizing customer reviews".

Proceedings of the ACM SIGKDD International Conference on

Knowledge Discovery & Data Mining (KDD-04), 2004.

Minqing Hu and Bing Liu. "Mining Opinion Features in Customer Reviews." Proceedings of Nineteeth National Conference on

Artificial Intellgience (AAAI-2004), 2004.