UA

EjerciciodeMejora

DanielAsensiRochDNI : **48776120C**

June 1, 2022

1 Introducción al problema

En este ejercicio se nos plantea la problemática de como implementaríamos en modelo clásico de Recuperación de Información pero introduciendo las técnicas del procesamiento de lenguajes naturales para que este alcance una mayor compresión del significado de los documentos. Además se nos indica que debemos teorizar el como incluiríamos el modelo de coseno con pesos.

2 Sistemas de recuperación de información y PLN

En la mayoría de los sistemas de recuperación la búsqueda de documentos relevantes a partir de una consulta depende casi exclusivamente de la presencia o ausencia de las mismas palabras en los documentos. Los modelos bayesianos han intentado mejorar los modelos clásicos basados en las proximidad entre espacios vectoriales.

Actualmente y en recientes estudios, se han utilizado en diferentes técnicas de PLN, pero los resultados no han sido muy satisfactorios. En el presente estudio que he realizado, para la aplicación de técnicas de PLN exitosas en modelos probabilísticos de RI, se han investigado las herramientas de **FreeLing** y **DeepPattern** dos herramientas de análisis robustas. Recordemos que muchas de las técnicas de las técnicas de PLN solo se encuentras disponibles para algunas lenguas.

Como planteamiento general **las técnicas de PLN permiten una ligera mejoría en la eficacia del sistema.**

3 Procesos lingüísticos estudiados

Para responder ante la pregunta que nos atañe en el correspondiente ejercicio, se han estudiado 11 sistemas referidos al modelo de análisis:

1. **Base**: tokenización y eliminación de stopwords
2. **lemas**: utilizar lematizador y tokenizador de Freeling y reducir palabras a lema flexivo
3. **lemas nominales, verbales y adjetivales**, además de eliminación de stopwords (lista preexistente)
4. **deps**: analizador sintáctico para identificación de núcleos y modificadores
5. **base + stem**: tokenización y eliminación de stopwords con stemming
6. **base + NPs**: tokens con nombres propios, simples y compuestos.
7. **base + lemas nominales, verbales y adjetivales**
8. **base + deps**
9. **base + lemas nominales, verbales y adjetivales + etiqueta morfosintáctica**
10. **base + NPs + deps**
11. **base + lemas nominales, verbales y adjetivales + tags + deps**

Las medidas de evaluación empleados miden el rendimiento de un sistema en función de su capacidad para devolver, por un lado, solo documentos relevantes, es decir la precisión y por otro lado tenemos los documentos relevantes, es decir la cobertura.

La precisión a los n documentos devueltos se calcula contando el número de documentos relevantes devueltos entre los n primeros documentos y dividiendo ese número por n, reflejando el rendimiento del sistema.

En la consecuente tabla se expresan los resultados obtenidos por diferentes test aplicados y recopilados por diferentes investigadores con los datos consecuentes, A y B colección de documentos anotados, t1 t2 y t3 lista de tópicos anotados y consultas cortas (c) medias (m) y largas (l).

	At1C	At1M	At1L	At2S	At2M	Bt3C	At3L	TOTAL
base	0.168	0.164	0.183	0.069	0.084	0.217	0.253	0.163
base+stem	0.16	0.163	0.172	0.066	0.083	0.23	0.26	0.162
lemas'	0.153	0.17	0.172	0.086	0.088	0.253	0.277	0.171
lemas	0.135	0.14	0.138	0.077	0.072	0.242	0.252	0.151
base+NPs	0.166	0.164	0.189	0.079	0.06	0.231	0.278	0.167
base+lemas'	0.185	0.163	0.178	0.086	0.093	0.25	0.271	0.175
base+lemas'+tags	0.182	0.16	0.173	0.079	0.083	0.244	0.266	0.170
deps	0.055	0.132	0.147	0.048	0.075	0.091	0.114	0.095
base+deps	0.173	**0.224**	**0.245**	0.087	**0.127**	0.219	0.273	**0.193**
base+NPs+deps	0.141	0.173	0.193	**0.092**	0.074	**0.254**	**0.317**	0.178
base+lemas'+tags+deps	**0.186**	0.187	0.203	0.09	0.102	0.246	0.282	**0.185**

Figure 1: Gráfica de precisiones obtenidas medias no interpoladas

4 Conclusiones

En los estudios realizados y tras estudiar los resultados reflejados en numerosos papers se destaca que el mejor rendimiento se obtiene al realizar las estrategias de **tokenización y extracción de dependencias** obteniendo una mejora del 19 por ciento ante las técnicas de RI tradicionales. Por el lado contrario el uso solo de **extracción de dependencias** disminuye en gran medida la precisión.

El uso de **stemming** no mejora los resultados obtenidos por las búsquedas RI base.

El uso de la **lematización**, excluyendo el tipo de la misma, iguala las preciones obtenidas mediante base y tokenización.

La inclusión de nombres propios, simples y compuestos **NPs + base** mejora la precisión en los primeros documentos relevantes, pero a medida que exploramos documentos vemos que esta precisión disminuye.

El uso de todas las técnicas disponibles no mejora en gran medida la precisión.

5 Mi aplicación

Ante los resultados estudiados y expuesto la aplicación que realizaría para la aplicación de técnicas de PLN en los modelos clásicos sería, realizar la lematización de absolutamente todos los términos de los documentos, esta práctica conllevaría un costo computacional elevado pero nos ayudaría en la eliminación de las stopwors ya que aquellas palabras que carezcan de lema serían eliminadas, ahorrándonos un análisis previo, acto seguido se procedería a la indexación de los términos, esto no solo se realizaría para los documentos sino también para las query de búsqueda.

Esta aplicación conllevaría fallos en la búsqueda de querys donde todos sus términos fueran stopwords como la siguiente: *"últimos empleos buenos"* o *"decir dos últimos trabajo"*.

Otro caso donde mi aplicación fallaría sería en el cual nuestra query o documento no tuviera lematización.

Por último aunque sea un fallo un tanto rebuscado mi aplicación fallaría ante idiomas donde no existan las lematizaciones de las palabras.

6 Aplicación de Modelo de coseno con pesos

La aplicación que le daría yo al modelo de coseno con pesos dentro del modelo clásico de RI sería teniendo en cuenta las posiciones de las palabras indexadas de la query en el documento dando más peso al documento que tenga los términos a distancias más parecidas a la de la query.