BENEMERITA UNIVERSIDAD AUTONOMA DE PUEBLA

Facultad de Ciencias de la Computación

Reporte del servicio social

Nombre del Programa:

Modelos para interpretar el tipo de similitud semántica textual entre pares de sentencias.

Community Question Answering

Número de Programa: 71279

Reporte de actividades

Prestador de Servicio Social:

Tapia Palacios Alberto

Asesor del Servicio social

Dra. Darnes Vilariño Ayala

Coordinadora de Servicio Social

M.C. Nelva Betzabel Espinoza Hernández

Contenido

Introducción

Los foros de respuestas a preguntas de una comunidad (Community Question Answering) están ganando popularidad en línea. Pero rara vez son moderados por un administrador, por lo cual tienen pocas restricciones sobre quién puede publicar y quién puede contestar una pregunta, de esta manera la mayoría de preguntas no llegan a ser contestadas en el tiempo esperado o bien se crean entradas duplicadas, es decir que la pregunta ya ha sido publicada anteriormente e incluso existe la posibilidad de estar respondida con comentarios útiles.

Los puntos positivos de estos foros son el orden en una discusión, es decir manejar adecuadamente todos los comentarios de los usuarios involucrados para mantener un orden específico, otro beneficio es que todos los usuarios registrados pueden participar en cualquier discusión y recibir notificaciones acerca de la misma, la discusión del foro puede ser sobre temas de interés general o específico, también se puede intercambiar información con usuarios de todas partes del mundo, así como la diversidad de comentarios y forma de pensar de los diferentes usuarios del foro, siendo un punto importante de este el que un usuario puede hacer cualquier pregunta libremente esperando respuestas que sean de utilidad y futuras referencias para otros usuarios. Sin embargo, los puntos negativos son la necesidad de una conexión a internet para poder participar en el foro, otro punto negativo es que no existe fiabilidad de los comentarios o inclusive pueden ser ofensivos, sin olvidar la necesidad de tiempo del administrador para revisar todas las preguntas con sus respectivas respuestas posibles y así relacionar las que sean más similares. No es raro que una pregunta tenga cientos de respuestas y viceversa, lo que hace que el usuario tenga la tarea de inspeccionar una por una y lograr encontrar su respuesta entre todas, esto no es una solución viable.

Justificación

La tarea tres de SemEval 2017 propone automatizar el proceso de encontrar las mejores respuestas en un foro de discusión creado por la comunidad, es decir que se recuperarán todas las preguntas que sean más similares a la proporcionada incluyendo sus respectivos comentarios ordenados por su nivel de utilidad "Buenos" (Good), "potencialmente útiles" (Potentially Useful) o "malos" (Bad) a partir de modelos la para interpretación del tipo de similitud semántica textual. Al automatizar este problema se logrará mejorar los resultados mostrados por parte del usuario pues estos le ayudarán a encontrar su información de una manera más eficiente.

Objetivos Generales

El principal objetivo del programa del servicio social es el de desarrollar un modelo para interpretar el tipo de similitud semántica textual entre pares de sentencias.

Objetivos específicos

- Extraer información necesaria de preguntas y respuestas de los archivos sin las etiquetas xml.
- Crear una colección de datos para buscar las relaciones que poseen las sentencias.
- Desarrollar un modelo que encuentre y clasifique sentencias mostrando las que son realmente útiles para el usuario.
- Comprobar la eficiencia del modelo desarrollado con las pruebas (test) y resultados (gold).

Descripción general del servicio

Crear algoritmos para el desarrollo de modelos que automaticen el proceso de encontrar y categorizar respuestas mostrando las que son realmente útiles para el usuario, es decir, que dada una pregunta y una colección de hilos de preguntas-respuestas creadas por la comunidad de usuarios(foro), se usarán para clasificar los puestos de comentarios que sean más apropiados para posibles respuestas a la pregunta proporcionada a partir de su similitud semántica, indexando y posteriormente buscando las relaciones que posee con la colección existente de preguntas o respuestas.

A continuación, se describe cada subtarea realizada para la se crearon distintos algoritmos para su solución (solo se trabajará con el conjunto de datos en inglés).

Subtarea A : Similaridad de Pregunta - Comentario

Proporcionando

- Una pregunta.
- Los primeros 10 comentarios del hilo de la pregunta.

Software

- Python

Herramientas y librerías utilizadas

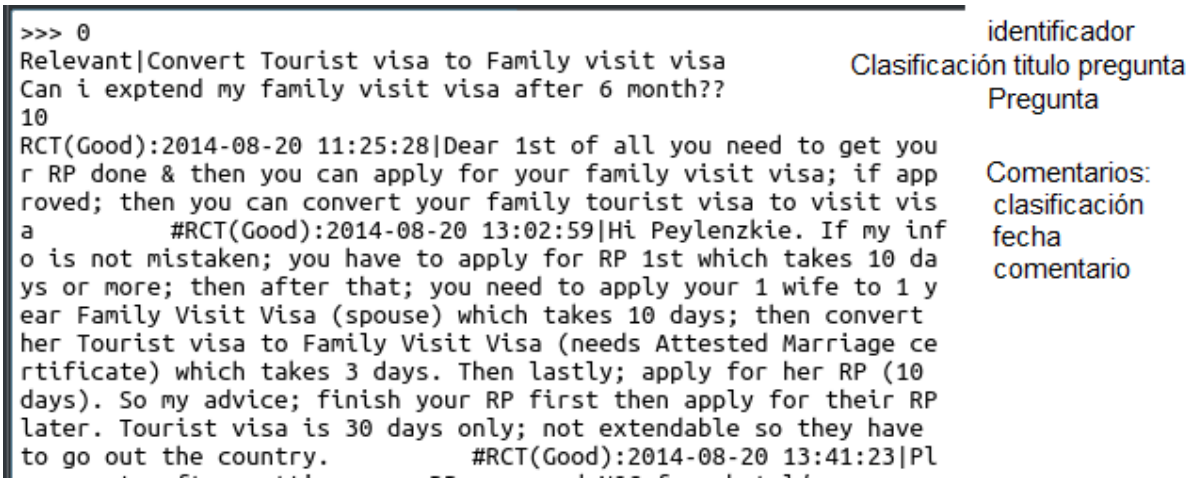
- xml.etree.ElementTree (Para lectura y desenlazar los subtipos del archivo xml)

- whoosh (Para crear la colección de datos indexando la similitud y relevancia, posteriormente categoriza el comentario dado a partir de una búsqueda)

Se organizaron estos 10 comentarios de acuerdo a su relevancia con respecto a la pregunta. Los comentarios Buenos serán clasificados por encima de los comentarios potencialmente útiles o malos; Los dos últimos no fueron distinguidos y son considerados "malos" para la evaluación.

Mostrando como salida la pregunta más similar a la proporcionada, con sus 10 respectivos comentarios ordenados por prioridad, para el caso de la pregunta “Can i exptend my family visit visa after 6 month?”.

A continuación se adjunta un fragmento de la primera respuesta para las pruebas realizadas, solamente como modelo de ejemplo.



Primer respuesta de la subtarea A.

El identificador de la pregunta sirve para visualizar cada respuesta más fácilmente, la clasificación título pregunta está conformada por la categoría que asignaron los usuarios (casos: PerfectMatch, Relevant, Irrelevant) con la parte del título de la pregunta en el foro, también se muestra la pregunta principal y los comentarios con su clasificación, fecha (día hora) y comentario de la pregunta.

Subtarea B : Pregunta-Pregunta Similaridad

Proporcionando

- Una nueva pregunta.
- Las primeras 10 preguntas relacionadas recuperado con Whoosh.

Software

- Python

Herramientas y librerías utilizadas

- xml.etree.ElementTree (Para lectura y desenlazar los subtipos del archivo xml)
- whoosh (Para crear la colección de datos indexando la similitud y relevancia, posteriormente categoriza el comentario dado a partir de una búsqueda)

Reclasificar las preguntas de acuerdo a su similitud con respecto a la pregunta original. En este caso, consideraremos las preguntas "PerfectMatch" y "Relevant" como buenas (considerando ambas "relevantes"), y deben ser clasificadas por encima de las preguntas "irrelevantes".

Para el caso de la pregunta “Who is the Best Hijacker here in QL”.

```
Pregunta a buscar:  Who is the Best Hijacker here in QL

total de resultados obtenidos en whoosh(MultifieldParser): 10

Lista de los 10 resultados principales

[similitud,Categoria,Pregunta(RELQ_RELEVANCE2ORGQ),PreguntaPrincipal]
[0.9595959595959596, 'PerfectMatch', 'What is the best way to hijack a thread?\t', u'Who is the Best Hijacker here in QL?']
[0.9595959595959596, 'PerfectMatch', 'Forum HiJackers\t', u'Who is the Best Hijacker here in QL?']
[0.9595959595959596, 'Relevant', 'Is QL now the rizks and pajju show ?\t', u'Who is the Best Hijacker here in QL?']
[0.9595959595959596, 'Relevant', 'Dirty trick indeed!\t', u'Who is the Best Hijacker here in QL?']
[0.9595959595959596, 'Irrelevant', 'Need a serious reply on this :)\t', u'Who is the Best Hijacker here in QL?']
[0.9595959595959596, 'Irrelevant', 'Lucky to have this kind of wife....\t', u'Who is the Best Hijacker here in QL?']
[0.9595959595959596, 'Irrelevant', 'I WANT TO GO CANADA\t', u'Who is the Best Hijacker here in QL?']
[0.9595959595959596, 'Irrelevant', 'FINALLY HERE :)\t', u'Who is the Best Hijacker here in QL?']
[0.9595959595959596, 'Irrelevant', '"thanx for QL admin!\t', u'Who is the Best Hijacker here in QL?']
[0.9595959595959596, 'Irrelevant', '"Boys & Girls Lets start posting ur good idea Regarding each other\t', u'Who is the Best Hijacker here in QL?']
```

Respuestas de la subtarea B.

Subtarea C: Pregunta-Comentario Externo Similaridad

Proporcionando

- Una nueva pregunta.
- Las primeras 10 preguntas relacionadas recuperado con Whoosh asociadas con sus primeros 10 comentarios que aparecen en su hilo.

Software

- Python

Herramientas y librerías utilizadas

- xml.etree.ElementTree (Para lectura y desenlazar los subtipos del archivo xml)
- whoosh (Para crear la colección de datos indexando la similitud y relevancia, posteriormente categoriza el comentario dado a partir de una búsqueda.)
- py_stringmatching (Obtención de similitud entre pares de sentencias)

Algoritmos utilizados:

- Soft TF:
  - TF: Term Frequency – El TF mide la frecuencia de uso de un término específico en una página o documento. Cuanto más largo sea un contenido mayor serán las veces que se use una keyword dentro del mismo.

- IDF: Inverse Document Frequency – IDF mide la importancia de un término específico por su relevancia dentro del documento. Dentro de esta métrica se excluyen las palabras vacías (Stopwords) como: “es”, “de”, “el/la” (términos que aparecerán de forma constante dentro de un documento o artículo, pero que realmente carecen de importancia).

Reclasificar los 100 comentarios (10 preguntas x 10 comentarios) de acuerdo a su relevancia con respecto a la pregunta original. Los comentarios "Bueno" se clasificaron por encima de los comentarios "potencialmente útiles" o "malos", que se considerarán malos para la evaluación.

Aunque se supone que los sistemas funcionan con 100 comentarios, tomamos una visión orientada a la aplicación en la evaluación: suponemos que a los usuarios potenciales se les presenta una lista relativamente corta de respuestas de candidatos (por ejemplo, 10 como en motores de búsqueda comunes hoy en día). Por lo tanto, a los usuarios les gustaría tener buenos comentarios para concentrarse en las primeras 10 posiciones, (es decir, todos los buenos comentarios clasificados antes de cualquier comentario no bueno).

Para el caso de la pregunta “Who is the Best Hijacker here in QL”.

```
Pregunta a buscar:  Who is the Best Hijacker here in QL

total de resultados obtenidos en whoosh(MultifieldParser): 10

Lista de los 10 resultados principales

Pregunta #1
similitud: 0.959595959596       PerfectMatch         PS: What is the best way to hijack a thread?
PP: Who is the Best Hijacker here in QL?
Comentarios:
1 RCT(Good):2008-06-18 14:38:14|i can refer you to a couple of specialists here...they are professional hija
ckers.. they reached their perfection in it.. 1. Ksarat16 - Managing Director 2. Smoke - Executive Project M
anager 3. DaRuDe - Head of Consultants Department and some other staff in this QL team... they can open a co
urse and will make a good money out of it!! Demand is there!
2 RCT(Bad):2008-06-18 14:30:00|...but i think i am doing it right now. do i???? ;-D
3 RCT(Bad):2008-06-18 14:30:47|how? chaud
4 RCT(Bad):2008-06-18 14:38:58|Ahem; ahem...
```

Primer respuesta de la subtarea C

Conclusión

Durante las actividades prestadas durante el servicio social se propuso un modelo para la solución de la tarea tres de SemEval 2017 utilizando la ayuda de herramientas y librerías desarrolladas en python, para poder clasificar las similitudes semánticas entre pares de oraciones (preguntas y comentarios del foro), todo el material de colección de datos de preguntas y respuestas mencionados fueron extraídos de la página oficial de semeval los cuales incluyen archivos de entrenamiento, pruebas y soluciones.

A trabajar con las subtareas A,B y C se logró hacer algoritmos creando un modelo que puede indexar el archivo de entrenamiento para formar una colección de datos, el cual fue utilizado para realizar búsquedas y categorizar las preguntas y comentarios del archivo de entrenamiento dependiendo a su vez de la similitud semántica que nos diera las pares de sentencias del archivo test. Este modelo propuesto originó un archivo de resultados los cuales a tener un formato similar al archivo gold(dado por SemEval), se compararon obteniendo resultados favorables, consiguiendo 1947 de aciertos de un total de 3270 comentarios, alcanzando un porcentaje de 59.54%, por lo tanto el modelo creado combinando los algoritmos realizados (subtareas), a tener un porcentaje mayor al 50% es aceptable debido a que se logró clasificar la mayoría de los comentarios efectivamente, identificando los que son útiles para el usuario ahorrándole tiempo y esfuerzo.

Sin embargo, estos resultados se podrían mejorar a incrementar la colección de datos, debido a que solo se incluyó un archivo de entrenamiento, por lo cual si se aumentaran daría resultados con una mayor precisión.

M.C. Nelva Betzabel Espinoza Hernández
COORDINADORA DE SERVICIO SOCIAL
FAC. DE CS. DE LA COMPUTACION

Dra. Darnes Vilariño Ayala
ASESORA DEL SERVICIO SOCIAL

Tapia Palacios Alberto