

Tarea 1 -Minería de datos

Rodrigo Fuenzalida

27 de Agosto 2014

1 Introduction

Para este trabajo se etiqueto manualmente 26 archivos correspondientes a 26 categorías conformadas por 30 tripletas <título, contenido, respuesta>.

2 Pregunta 1

Si en una página de respuestas, es decir, una tripleta <título, contenido, respuesta> hay discrepancias para la clase dada manualmente a una palabra determinada.

2.1 ¿Qué observa?

Dado que se tiene conocimiento de la categoría a la cual pertenece una tripleta es mucho más fácil determinar la clase de una palabra en específico, puesto que el contexto que entrega la categoría de la tripleta es muy importante, y tiene directa influencia en la decisión de categorías algo como organización, locación o persona.

2.2 ¿Qué aprende de esto?

Qué es importante tener conocimiento previo del tipo de contenido de una tripleta, el tener conocimiento de la categoría a la cual pertenece una tripleta es de muchísima ayuda para clasificar, puesto que nos entrega un parámetro más para tener en consideración al momento de hacer una elección.

2.3 ¿Hay una clase más susceptible a las discrepancias?

Ciertamente, las clases como locaciones o organizaciones muchas veces se ven complicadas por el tipo de nombre y uso que se les da, por ejemplo en un contexto en el cual se están buscando restaurantes para salir a comer en la noche, podemos encontrar un sin fin de lugares que también representan organizaciones.

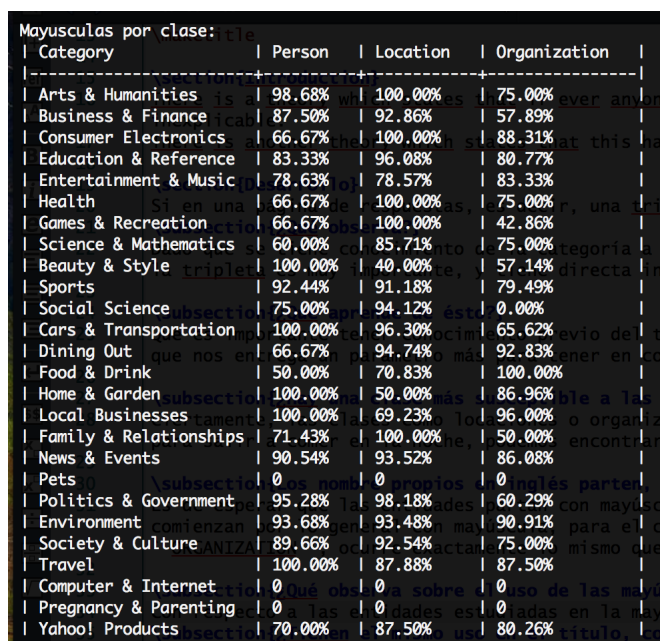
3 Pregunta 2

3.1 Los nombre propios en inglés parten, en su gran mayoría, con mayúsculas. Es esperable que las entidades estudiadas en esta tarea también partan con mayúsculas.

Es de esperar que las entidades partan con mayúsculas ya que las tres clases que estamos analizando son nombres propios, en el caso de “PERSON”, nos estamos refiriendo a personas cuyos nombres comienzan por lo general con mayúscula, para el caso de las “LOCATION”, pasa lo mismo, al tratarse de localidades, se utiliza mayúscula al inicio de cada palabra que la compone. Con respecto a “ORGANIZATION”, ocurre exactamente lo mismo que con las otras entidades.

3.2 ¿Qué observa sobre el uso de las mayúsculas en Yahoo! Answers?

Con respecto a las entidades estudiadas en la mayoría de las ocasiones el uso de las mayúsculas es bien utilizado, sobre todo cuando se presentan nombres de personas, lugares y organizaciones.



| Category | Person | Location | Organization |
|------------------------|---------|----------|--------------|
| Arts & Humanities | 98.68% | 100.00% | 75.00% |
| Business & Finance | 87.50% | 92.86% | 57.89% |
| Consumer Electronics | 66.67% | 100.00% | 88.31% |
| Education & Reference | 83.33% | 96.08% | 80.77% |
| Entertainment & Music | 78.63% | 78.57% | 83.33% |
| Health | 66.67% | 100.00% | 75.00% |
| Games & Recreation | 66.67% | 100.00% | 42.86% |
| Science & Mathematics | 60.00% | 85.71% | 75.00% |
| Beauty & Style | 100.00% | 40.00% | 57.14% |
| Sports | 92.44% | 91.18% | 79.49% |
| Social Science | 75.00% | 94.12% | 0.00% |
| Cars & Transportation | 100.00% | 96.30% | 65.62% |
| Dining Out | 66.67% | 94.74% | 92.83% |
| Food & Drink | 50.00% | 70.83% | 100.00% |
| Home & Garden | 100.00% | 50.00% | 86.96% |
| Local Businesses | 100.00% | 69.23% | 96.00% |
| Family & Relationships | 71.43% | 100.00% | 50.00% |
| News & Events | 90.54% | 93.52% | 86.08% |
| Pets | 0 | 0 | 0 |
| Politics & Government | 95.28% | 98.18% | 60.29% |
| Environment | 93.68% | 93.48% | 90.91% |
| Society & Culture | 89.66% | 92.54% | 75.00% |
| Travel | 100.00% | 87.88% | 87.50% |
| Computer & Internet | 0 | 0 | 0 |
| Pregnancy & Parenting | 0 | 0 | 0 |
| Yahoo! Products | 70.00% | 87.50% | 80.26% |

Figure 1: Mayúsculas por clase

En la Figura 1, se aprecian las cantidades en porcentaje del uso de mayúsculas en los name entities obtenidos mediante el etiquetado manual, si bien hay categorías que tienen un porcentaje de cero, esto es debido a que el uso en esas

categorías no es tan relevante para el tipo de tripleta que se está presentando. ¿Tienen el mismo uso en el título, contenido y respuestas? Si, esto se aprecia con mayor claridad en preguntas que tienen relación con persona, lugares y organizaciones, puesto que la mayoría de éstas se escriben con mayúscula inicial.

3.3 ¿Qué puede corroborar acerca del uso de la mayúscula y los “tokens”?

En este caso, nos sirve de ayuda para poder discriminar un token de un name entity manera más fácil, puesto que en las tres clases utilizadas para etiquetar un patrón común es el uso de mayúsculas al inicio de cada palabra, y desde luego al ser éstos nombres propios.

3.4 De las palabras que comienzan con mayúsculas ¿cuántas fueron reconocidas como “named entities” y cuántas como “tokens”?

En la Figura 2, se observa la cuenta de la cantidad de palabras que son iniciadas con mayúsculas para cada “name entity” reconocido, a su vez se presentan las palabras que no comienzan con mayúsculas, representando parte importante de los datos como “tokens”, también se observa la cuenta de cuantos “Tokens” comienzan con mayúscula.

| Categoria | Cantidad |
|--------------|----------|
| Location | 235 |
| Organization | 355 |
| Person | 515 |
| Tokens | 4679 |
| Resto | 14185 |

Figure 2: Cuenta de palabras iniciadas con mayúscula

4 Pregunta 3

4.1 ¿Cuántas de las palabras que fueron sugeridas en una de las tres clases de “named entity” y no fueron elegidas por ud. como tal fueron confusiones producto de errores ortográficos?

Para este caso ningún “named entity” fue seleccionado por tener una discrepancia por faltas ortográficas, más bien los “named entity” que fueron sugeridos erradamente, fueron sugeridos mal por estar en un contexto errado.

4.2 ¿Qué puede concluir de aquello?

Las sugerencias son muy útiles al momento de tener que determinar la pertenencia de una palabra a una clase determinada. Aportan mayor información al contexto, lo que permite una clasificación manual más rápida y segura.

5 Pregunta 4

5.1 ¿Qué sucede con las estructuras como tablas y listas? ¿Facilitan o dificultan la etiquetación?

Facilitan mucho el etiquetado, ya que por lo general estas contienen a los “named entity”, y su vez, estos tienen directa relación con la pregunta realizada.

5.2 ¿Qué sucede cuando la respuesta es la entidad?

En este caso depende mucho de qué tan bien explicada esté la pregunta, ya que si no se explica bien el contexto, es muy difícil derivar una asociación coherente con la entidad.

5.3 ¿En este último caso se obtiene una buena o mala identificación? Ejemplifique. En estos casos ¿Las sugerencias son o no un aporte?

Por lo general las sugerencias son un gran aporte ya que nos dan un indicio de que criterio utilizar al momento de etiquetar una entidad.

6 Pregunta 5

6.1 Utilice MontyLingua para hacer un análisis morfológico y sintáctico de las palabras en las tripletas. ¿Qué observa?

En general las entidades no cambian su forma, los cambios más notorios son en aquellas entidades que podrían representar un plural, en este caso MontyLingua las lleva a su forma singular, lo cual en cierto sentido no alteraría el objetivo de la entidad. Por otra parte muchos token son eliminados ya que no representan mayor valor sintáctico dentro del conjunto de palabras que forman una tripleta.

6.2 ¿Las palabras que ud. etiqueto como entidades son susceptibles a ser mapeadas a una raíz o éstas no cambian?

En su mayoría no cambian, salvo excepciones como se puede ver en la Figura 3.

| | |
|--------------|-------------|
| Entidad | Lema |
| Dreamworks | Dreamwork |
| Dreamworks | Dreamwork |
| Gottfried | Gottfry |
| Euripides | Euripide |
| Jean-Jacques | jean-jacque |
| Jones | Jone |
| Walters | Walter |
| Carla | Carlum |
| DALLAS | DALLA |
| US | U |
| Windows | Window |
| Windows | Window |
| AVS | AV |
| Motorola | Motorolum |
| US | U |
| Hughes | Hugh |
| Isaacs | Isaac |
| Isaacs | Isaac |
| US | U |
| Bowling | Bowl |

Figure 3: Entidades lematizadas

6.3 ¿Hay una clase más afectada?

En la Figura 4 se puede ver como la clase “organization” es la que sufre mayores cambios con el uso de MontyLingua.

| | | |
|-----------------------------|----------|--------------|
| Clases con mayores cambios: | | |
| Person | Location | Organization |
| 43 | 58 | 102 |

Figure 4: Entidades con cambios

6.4 ¿Qué sucede con los tokens? En cuanto a las categorías sintáctica, ¿Cuál es la distribución en el conjunto de datos?

Como se puede ver en la Figura 5 las categorías sintácticas más relevantes dentro de los “tokens” es la clase NN, que representa a los sustantivos, y esto nos indica que están mayormente presentes en el texto que estamos analizando.

| Categoría | Cuenta |
|-----------|--------|
| NN | 33914 |
| IN | 20870 |
| PRP | 17205 |
| DT | 16983 |
| JJ | 12617 |
| RB | 12064 |
| VB | 11896 |
| VBP | 8718 |
| NNS | 8554 |
| CC | 7897 |
| NNP | 7229 |
| VBZ | 6505 |
| TO | 5354 |
| CD | 4878 |
| MD | 4204 |
| VBG | 4028 |
| PRP\$ | 3977 |
| VBD | 3923 |
| VBN | 3699 |
| WRB | 1609 |
| SVM | 1602 |
| WP | 1204 |
| WDT | 850 |
| JJR | 831 |
| EX | 559 |
| POS | 447 |
| JJS | 429 |
| UH | 394 |
| FW | 236 |
| NNPS | 216 |
| RBR | 211 |
| RBS | 140 |
| RP | 118 |
| PDT | 32 |
| WP\$ | 2 |

Figure 5: Entidades con cambios

6.5 ¿Cuáles son las cinco clases sintácticas más prominentes en cada una de las 26 categorías?

Esto se puede ver desde la Figura 6 hasta la Figura 31.

| Categoria: Arts & Humanities | |
|------------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1324 |
| IN | 944 |
| DT | 797 |
| PRP | 726 |
| JJ | 566 |

Figure 6: Arts & Humanities

| Categoria: Business & Finance | |
|-------------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1380 |
| IN | 790 |
| PRP | 676 |
| DT | 610 |
| VB | 465 |

Figure 7: Bussiness & Finance

| Categoria: Consumer Electronics | |
|---------------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 882 |
| IN | 410 |
| DT | 377 |
| PRP | 342 |
| NNP | 288 |

Figure 8: Consumer Electronics

| Categoria: Education & Reference | |
|----------------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 975 |
| IN | 621 |
| DT | 520 |
| PRP | 503 |
| JJ | 416 |

Figure 9: Education & Reference

| Categoria: Entertainment & Music | |
|----------------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 951 |
| DT | 489 |
| IN | 486 |
| PRP | 457 |
| JJ | 329 |

Figure 10: Entertainment & Music

6.6 ¿Cuál(es) es la categoría sintáctica más frecuente entre las palabras que etiqueto manualmente como “named entity”?

En la Figura 32 se puede ver que la categoría sintáctica NNP es la más común dentro de los “named entity” clasificados.

| Categoria: Health | |
|----------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1463 |
| IN | 842 |
| PRP | 759 |
| DT | 589 |
| JJ | 559 |

Figure 11: Healh

| Categoria: Games & Recreation | |
|-------------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 932 |
| IN | 585 |
| PRP | 485 |
| DT | 460 |
| NNP | 391 |

Figure 12: Games & Recreation

| Categoria: Science & Mathematics | |
|----------------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1695 |
| CD | 1086 |
| DT | 601 |
| IN | 581 |
| SYM | 476 |

Figure 13: Science & Mathematics

| Categoria: Beauty & Style | |
|---------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1083 |
| PRP | 625 |
| IN | 602 |
| JJ | 538 |
| DT | 455 |

Figure 14: Beauty & Style

| Categoria: Sports | |
|----------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 880 |
| IN | 571 |
| DT | 508 |
| PRP | 499 |
| NNP | 374 |

Figure 15: Sports

| Categoria: Social Science | |
|---------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1476 |
| PRP | 1208 |
| IN | 1194 |
| DT | 791 |
| RB | 752 |

Figure 16: Social Science

| Categoria: Cars & Transportation | |
|----------------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1399 |
| IN | 833 |
| DT | 817 |
| PRP | 568 |
| RB | 480 |

Figure 17: Cars & Transportation

| Categoria: Dining Out | |
|-----------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 973 |
| IN | 623 |
| DT | 507 |
| PRP | 452 |
| JJ | 419 |

Figure 18: Dining Out

| Categoria: Food & Drink | |
|-------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1011 |
| IN | 551 |
| PRP | 426 |
| JJ | 413 |
| DT | 387 |

Figure 19: Food & Drink

| Categoria: Home & Garden | |
|--------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1270 |
| IN | 693 |
| DT | 635 |
| PRP | 595 |
| VB | 485 |

Figure 20: Home & Garden

| Categoria: Local Businesses | |
|-----------------------------|--------|
| Categoria Sintactica | Cuenta |
| ----- | ----- |
| NN | 928 |
| IN | 493 |
| PRP | 319 |
| DT | 306 |
| NNP | 251 |

Figure 21: Local Businesses

| Categoria: Family & Relationships | |
|-----------------------------------|--------|
| Categoria Sintactica | Cuenta |
| ----- | ----- |
| PRP | 1856 |
| NN | 1537 |
| IN | 1376 |
| RB | 1042 |
| VB | 875 |

Figure 22: Family & Relationships

| Categoria: News & Events | |
|--------------------------|--------|
| Categoria Sintactica | Cuenta |
| ----- | ----- |
| NN | 1117 |
| IN | 891 |
| DT | 773 |
| PRP | 584 |
| JJ | 578 |

Figure 23: News & Events

| Categoria: Pets | |
|----------------------|--------|
| Categoria Sintactica | Cuenta |
| ----- | ----- |
| NN | 1531 |
| IN | 912 |
| PRP | 784 |
| DT | 769 |
| JJ | 634 |

Figure 24: Pets

| Categoria: Politics & Government | |
|----------------------------------|--------|
| Categoria Sintactica | Cuenta |
| ----- | ----- |
| NN | 1700 |
| IN | 1119 |
| DT | 1051 |
| PRP | 963 |
| VB | 650 |

Figure 25: Politics & Government

| Categoria: Environment | |
|------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 2773 |
| IN | 1674 |
| DT | 1416 |
| JJ | 1052 |
| PRP | 941 |

Figure 26: Environment

| Categoria: Society & Culture | |
|------------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1027 |
| IN | 1394 |
| DT | 1109 |
| PRP | 1000 |
| JJ | 865 |

Figure 27: Society & Culture

| Categoria: Travel | |
|----------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1201 |
| IN | 948 |
| DT | 703 |
| PRP | 540 |
| RB | 464 |

Figure 28: Travel

| Categoria: Computer & Internet | |
|--------------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1161 |
| IN | 531 |
| DT | 477 |
| NNP | 402 |
| PRP | 400 |

Figure 29: Computer & Internet

| Categoria: Pregnancy & Parenting | |
|----------------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1170 |
| PRP | 841 |
| IN | 824 |
| DT | 570 |
| RB | 560 |

Figure 30: Pregnancy & Parenting

| Categoria: Yahoo! Products | |
|----------------------------|--------|
| Categoria Sintactica | Cuenta |
| NN | 1195 |
| SVM | 611 |
| IN | 472 |
| DT | 420 |
| PRP | 402 |

Figure 31: Yahoo! Products

| Categoría Sintáctica | Cuenta |
|----------------------|--------|
| NNP | 2119 |
| NN | 305 |
| JJ | 29 |
| PRP | 25 |
| NNS | 22 |

Figure 32: Categoría sintáctica para “named entity”

7 Pregunta 6

7.1 Antes de una “named entity”: ¿Cuál es la clase sintáctica más común?, y después de una “named entity” ¿Qué sucede?

La Figura 33, muestra el conteo de las clases sintácticas que se detectan antes de un “named entity”. Se puede apreciar como las clases sintácticas varían de acuerdo a si están antes o después de una “named entity”, esto debido a la composición del texto, esto implica que algunas clases sintácticas aparecerían más que el resto.

| Categoría Sintáctica | Cuenta |
|----------------------|--------|
| NNP : | 544 |
| IN | 499 |
| DT | 234 |
| CC | 142 |
| NN | 69 |

Figure 33: Clases sintácticas antes.

| Categoría: Yahoo! Products | |
|----------------------------|--------|
| Categoría Sintáctica | Cuenta |
| NNP : | 532 |
| CC | 171 |
| NN | 157 |
| VBZ | 149 |
| POS | 118 |

Figure 34: Clases sintácticas después.

7.2 ¿Es para las tres clases de entidades lo mismo?

En la Figura 35, se muestran los resultados obtenidos para cada una de las entidades. Se puede ver que para cada una de las entidades hay clases sintácticas

que predominan más que otras.

| Organization: | |
|----------------------|--------|
| Categoría Sintáctica | Cuenta |
| ----- | ----- |
| NNP : | 170 |
| IN | 136 |
| DT | 118 |
| CC | 35 |
| NN | 27 |

| Location: | |
|----------------------|--------|
| Categoría Sintáctica | Cuenta |
| ----- | ----- |
| IN | 251 |
| NNP : | 84 |
| DT | 83 |
| CC | 48 |
| TO | 34 |

| Person: | |
|----------------------|--------|
| Categoría Sintáctica | Cuenta |
| ----- | ----- |
| NNP : | 290 |
| IN | 112 |
| CC | 59 |
| NNP | 34 |
| DT | 33 |

Figure 35: Clases sintácticas por entidad

8 Pregunta 7

- 8.1 Considerando un ejercicio binario 0/1, si una palabra es o no entidad. Calcule la accuracy, la precisión, el recall y el F-Score de la clase positiva (entidad). Para ésto, considere las sugerencias como respuesta automática y sus etiquetas manuales como la “verdad absoluta”? También calcule la entropía del conjunto de datos considerando 0/1.

La Figura 36 muestra los resultados obtenidos, tanto para accuracy, precision, recall y F-Score, también se presentan los resultados para la Entropía de nuestro conjunto de datos. En este caso el accuracy está cerca del 70% lo que nos indica que al rededor de un 70% de las sugerencias fueron tomadas en cuenta al momento de etiquetar.

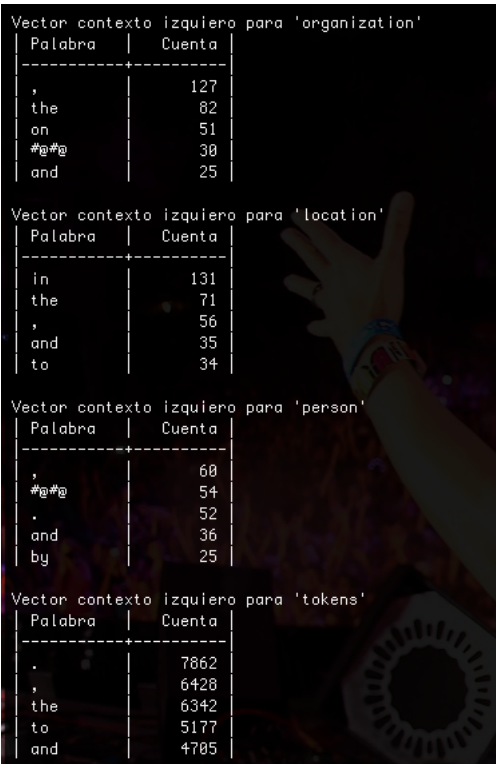
```
Accuracy: 0.78  
Precision: 0.72  
Recall: 0.74  
F-Score: 0.73  
  
Entidades totales: 2571  
Tokens totales: 222642  
Palabras totales: 225213  
Entropía: 0.09
```

Figure 36: Calculo accuracy, precisión, recall, F-Score, Entropía.

9 Pregunta 8

9.1 ¿Qué observa para los ocho vectores?

La Figura 37, muestra el resultado obtenido para los vectores de contexto izquierdo en el que podemos apreciar y de terminar que para este caso la palabra que más veces aparece antes de una de las clases es “,” por otra parte palabras como “and”, “the”, “is”, “to”, tienen alta frecuencia de aparición. En la Figura 38 se muestran los resultados para las palabras con mayor aparición para cada clase. Se utilizó el carácter #@#@ se utilizó como cominzo y fin de linea.



| Palabra | Cuenta |
|---------|--------|
| , | 127 |
| the | 82 |
| on | 51 |
| #@#@ | 30 |
| and | 25 |

| Palabra | Cuenta |
|---------|--------|
| in | 131 |
| the | 71 |
| , | 56 |
| and | 35 |
| to | 34 |

| Palabra | Cuenta |
|---------|--------|
| , | 60 |
| #@#@ | 54 |
| . | 52 |
| and | 36 |
| by | 25 |

| Palabra | Cuenta |
|---------|--------|
| . | 7862 |
| , | 6428 |
| the | 6342 |
| to | 5177 |
| and | 4705 |

Figure 37: Vector de contexto izquierdo

10 Conclusion

10.1 ¿Qué indicadores hay en el texto que podría indicarme si una palabra es una “named entity” o no, si su clase respectiva?

Como pudimos apreciar a través de este trabajo, existen muchos patrones los cuales no sirven para poder discriminar una “named entity”. Por ejemplo, tenemos las categorías sintácticas que ya nos dan un indicio de lo que puede ser el contexto del texto que se está analizando o la presencia de algunos atributos importantes o la estructura de las respuesta, entregando así mayor información y facilitando la labor de etiquetado manual.

10.2 ¿Qué conclusiones cree que también se darían para el caso del idioma español?

Posiblemente dependería mucho de las reglas que se utilicen para clasificar, como el español es un lenguaje rico en reglas quizás esto sea un punto a favor

| Vector contexto derecho para 'organization' | |
|---|--------|
| Palabra | Cuenta |
| , | 115 |
| 's | 68 |
| and | 48 |
| . | 34 |
| is | 23 |

| Vector contexto derecho para 'location' | |
|---|--------|
| Palabra | Cuenta |
| , | 121 |
| . | 59 |
| and | 58 |
| ? | 35 |
| is | 23 |

| Vector contexto derecho para 'person' | |
|---------------------------------------|--------|
| Palabra | Cuenta |
| , | 69 |
| . | 56 |
| and | 32 |
| is | 31 |
| 's | 31 |

| Vector contexto derecho para 'tokens' | |
|---------------------------------------|--------|
| Palabra | Cuenta |
| . | 9618 |
| the | 6469 |
| , | 6375 |
| to | 5227 |
| and | 4669 |

Figure 38: Vector de contexto derecho

al momento de buscar la clasificación de las distintas entidades. Ahora bien, esto también puede llegar a ser un problema ya que si no se tiene conocimiento previo de estas reglas puede llegar a ser muy difícil clasificar.

10.3 ¿Qué características piensa ud. harían más fácil/difícil el reconocimiento de entidades en español?

El genero, los tiempos verbales, las preposiciones, etc..., todas estas características hacen que el etiquetar en español sea algo más complicado, sobre todos para aquel que no domina el lenguaje o no tiene conocimiento completo de las reglas gramaticales con las que cuenta el español.

10.4 ¿Estaría un banco interesado en reconocer entidades? ¿En qué tipo de casos y para qué?

Claramente que si, sería interesante el poder trabajar con las descripciones de los movimientos de las personas, con el objetivo de poder identificar que compran

más o donde se va su dinero, así los bancos podrían separar a los grupos de personas y poder recomendarles productos acordes a su clase de gastos.

10.5 ¿Al gobierno le interesaría identificar entidades en alguna de sus base de datos?

Como está de moda el emprender y el gobierno a hecho algunas plataformas digitales para poder facilitar eso, sería interesante poder trabajar en el ámbito del registro de patentes o marcas donde el poder identificar entidades como: una “organización” o la persona que esté a cargo del proyecto, puede ser de mucha ayuda para incluso poder clasificar a que rubro pertenece dicho emprendimiento y si hacer un mejor análisis de cuál es la tendencia real, al momento de crear una empresa.