|  |
| --- |
|  |
| Informe Tarea 2 |
| Tecnologías de búsqueda en la web |
|  |
| **Leonardo Santis Arenas**  **Alonso Orellana Abarca**  **Rodrigo Martínez Díaz** |
| **02/06/2015** |

|  |
| --- |
|  |

# Resultados obtenidos

Con la utilización de Mallet, se proceso la colección trabajada en la tarea 1 de títulos y abstracts. Se crearon varios topic models utilizando el comando train-topics bajo los input.mallet generados, ajustando los parámetros α, β, k. Se guardan los doc-topics y los topic-word-weights de cada topic model, los cuales fueron estudiados. Las configuraciones de parámetros utilizados para las tres colecciones (abstracts, titles y both), fueron las siguientes:

|  |  |  |
| --- | --- | --- |
| α | β | k |
| 10 | 0,04 | 7 |
| 20 | 0,05 | 10 |
| 25 | 0,03 | 3 |
| 30 | 0,02 | 8 |
| 35 | 0,0015 | 6 |
| 50 | 0,01 | 5 |

Mediante una implementación en Python, se calcula la verosimilitud asociada de cada topic model, utilizando la siguiente formula:

Los resultados encontrados, para cada topic model, fueron los siguientes:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| α | β | k | Abstract | Title | Both |
| 10 | 0,04 | 7 | 143790,80516 | 107336,26682 | 137930,68457 |
| 20 | 0,05 | 10 | 189810,00188 | 137387,71967 | 181790,65561 |
| 25 | 0,03 | 3 | 78200,92963 | 60596,68609 | 75612,79990 |
| 30 | 0,02 | 8 | 163913,72279 | 119129,65536 | 157628,93636 |
| 35 | 0,015 | 6 | 132335,55019 | 98694,72974 | 128396,21424 |
| 50 | 0,01 | 5 | 116318,73146 | 87229,97067 | 112594,75529 |

Para cada uno de los tres inputs, la mejor verosimilitud se observo en la configuración de parámetros α=25, β=0,03, k=3. Una menor cantidad de tópicos (k) reduce la verosimilitud observada, podemos ver que cada para cada uno de los tres inputs, k es directamente proporcional a la verosimilitud observada, la configuración de los parámetros α y β se ve opacada por la relevancia de la cantidad de tópicos. Por lo que no es posible inferir nada al respecto, los gráficos se pueden encontrar en el anexo.

Las top-keywords para las mejores configuraciones encontradas (α=25, β=0,03, k=3), son las siguientes:

Abstracts:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Topic 1 |  | Topic 2 |  | Topic 3 |  |
| C | 4989 | N | 929 | Equation | 374 |
| T | 4465 | Space | 583 | Problem | 356 |
| Mallet | 3768 | K | 467 | Function | 326 |
| Sq | 3759 | Group | 446 | Two | 319 |
| Xw | 3747 | prove | 392 | one | 314 |

Titles:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Topic 1 |  | Topic 2 |  | Topic 3 |  |
| X | 154 | Group | 140 | Equation | 136 |
| Space | 152 | Function | 132 | Problema | 98 |
| System | 108 | Algebra | 107 | Field | 92 |
| Non | 73 | Operator | 103 | Quantum | 90 |
| xe | 59 | finite | 97 | theory | 88 |

Both:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Topic 1 |  | Topic 2 |  | Topic 3 |  |
| Time | 351 | N | 1194 | Space | 735 |
| Model | 337 | X | 714 | Equation | 510 |
| Two | 325 | K | 638 | Theory | 415 |
| System | 298 | Group | 587 | Field | 382 |
| Random | 291 | p | 529 | operator | 354 |

Se puede observar la existencia de tres tópicos. En los resultados conjuntos (Both): Topic 1 pareciera hablar de sistemas y modelamiento, Topic 2 acerca de variables algebraicas o físicas, Topic 3 acerca del espacio (pudiendo ser matemático o físico) y sus teorías/ecuaciones.

# Anexo

Relación entre cantidad de tópicos k y verosimilitud observada en cada uno de los tres inputs:

Relación entre parámetro β y verosimilitud observada en cada uno de los tres inputs:

Relación entre parámetro α y verosimilitud observada en cada uno de los tres inputs: