TP Text Mining

**Base de dato:**

1. *Cambiar el dataset que alimenta el LDA:*

* ~~Que tome todos los documentos desde 2016 para atrás~~
* ~~Desde 2017 en adelante que tome 1810 documentos por año para que sea proporcional y no le de mayor peso a los topics de los últimos años~~

1. ~~Limpiar script LDA. Vamos a usar este de base~~
2. *Para comparar por año:*

Calcular las distribuciones de los topics del punto 1 por año de la siguiente manera:

Hasta 2010 por década y luego en adelante por año para tener cual es la distribución por año. (Probar ventanas)

4) Iterar 50 topics y épocas y ver que sale

1-4 [mundodie2@hotmail.com](mailto:mundodie2@hotmail.com)[Mateosuster@gmail.com](mailto:Mateosuster@gmail.com)

5) Utilizar gensim dynamic que toma como parámetro el año para calcular la evaluación de cambio de topics:[https://radimrehurek.com/gensim/models/ldaseqmodel.html#](https://radimrehurek.com/gensim/models/ldaseqmodel.html) [Javgamboa@gmail.com](mailto:Javgamboa@gmail.com)

I**deas:**

* Ver de fijar algunos topics fijo y alimentar el LDA (con algo que no sea de inicio random)
* Comparar los topics solo tomando los workshops vs los principales

**Género** [Clara Trinco](mailto:claratrinco@gmail.com)

Dividir con el primer nombre y ver la evolución de proporción hombre/mujer

Ver si los topics cambiaron

Vamos a usar dos librerias en principio

* getgender de guender guesse <https://pypi.org/project/gender-guesser/>
* NTKL <https://www.geeksforgeeks.org/python-gender-identification-by-name-using-nltk/>

Pasos\_ entender los unknow y ver si hay alguna libreria que tenga nombres en Indi u otros lenguajes ya que tenemos ⅓ mas o menos que no los estamos encontrando.

Otras notas

topicos de sklearn (ver imagen guardada)

1 embedings

6 sequence 2 sequence

3 herramientas - implementacion

7 palabras en otro idioma? articulos? parte de nombres?

9 emotions analysis (hater etc)

5 computation and social sceince,