En el contexto de la genética de poblaciones , JK Pritchard , M. Stephens y P. Donnelly propusieron LDA en 2000. En el contexto del aprendizaje automático , donde se aplica más ampliamente hoy, LDA fue redescubierto independientemente por David Blei , Andrew Ng y Michael I. Jordan en 2003, y presentado como un modelo gráfico para el descubrimiento de temas. A partir de 2019, estos dos documentos tenían 24,620 y 26,320 citas respectivamente, lo que los ubica entre los más citados en los campos de aprendizaje automático e inteligencia artificial..

Natural language processing
Todo el procesamiento de datos de texto lo tomamos desde machine learning with pyspark: whit Natural Language Processing and Recommender Systems de Pramod Singh.

Reading the corpus
El corpus es colección completa de documentos desde los cuales vamos a procesar y analizar nuestros datos.

Tokenization
La tokenización es método por el cual dividimos los datos que se encuentran en la colección para así procesarlos de manera más fácil.

Stopword
Lo que nos permite este método es quitar las palabras de poco valor en el texto, como lo son los conectores o palabras cortas que nos nos ayudan mucho en la analitica del mismo.

singh, P.{2019}. Machine learning with PySpark: with Natural language Processing and Recommender Systems. Recuperado de https://doi.org/10.1007/978-1-4842-4131-8

Pritchard, JK; Stephens, M .; Donnelly, P. (junio de 2000). "Inferencia de la estructura de la población utilizando datos de genotipo multilocus" . *La genética* . **155** (2): *págs.* 945–959. ISSN 0016-6731 . PMID 10835412 .