TalleR avanzado
Text Mining-NLP
Topic Mining

Pedro Concejero

pedro.concejero@u-tad.com

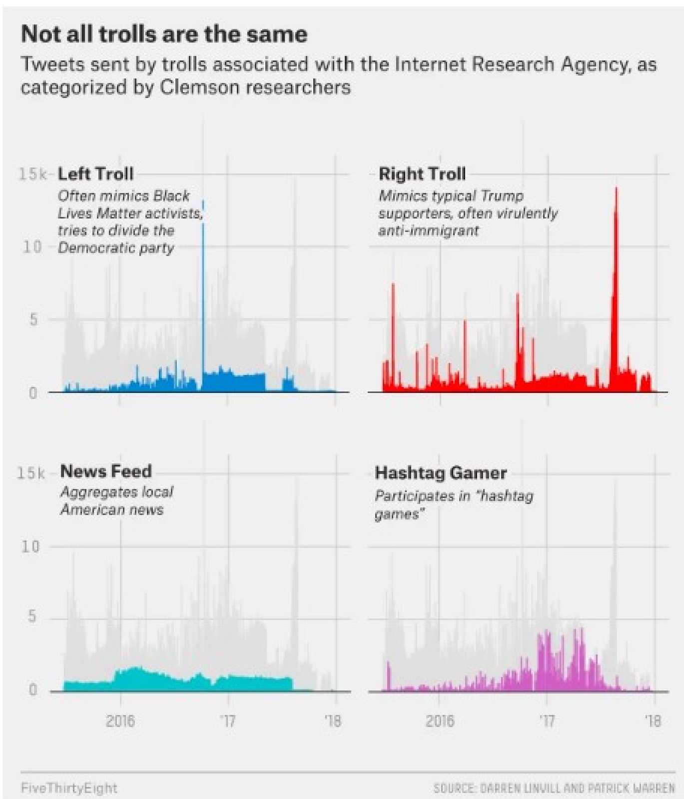
Enlaces para el talleR

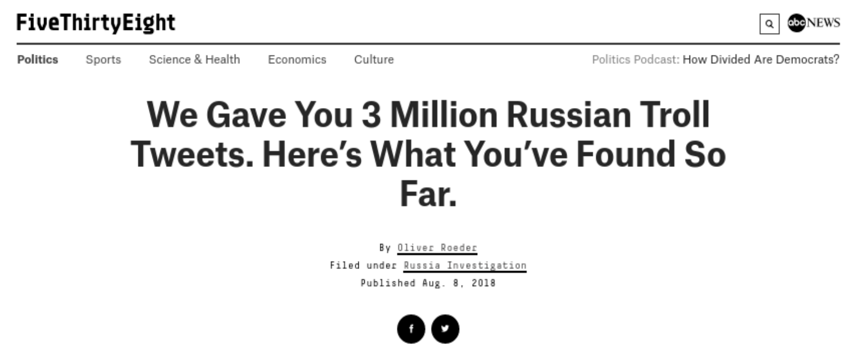
Repo github pedroconcejero/russian_trolls_topicmining:

https://github.com/pedroconcejero/russian_trolls_topicmining

Topic mining sobre 3 millones de troll-tweets

https://fivethirtyeight.com/features/what-you-found-in-3-million-russian-troll-tweets/





	0	1
Fearmonger	9368	1388
HashtagGamer	60383	154054
LeftTroll	70137	320932
NewsFeed	585885	1050
NonEnglish	363871	403894
RightTroll	345883	250975
Unknown	10106	2825
Commercial	113693	7872

quanteda

https://quanteda.io/

Keyword in context

```
   [text73, 1]                                   | Obama |
  [text83, 19]                to the USA oh sorry | Obama |
   [text87, 1]                                   | Obama |
 [text325, 12]             Bob Marley singing' No | Obama |
  [text351, 9]     Pope definitely should talk with | Obama |
  [text375, 1]                                   | Obama |
  [text544, 1]                                   | Obama |
  [text572, 1]                                   | Obama |
  [text589, 1]                                   | Obama |
  [text800, 6]           '@thehill have Nike payed | Obama |
  [text834, 2]                          President | Obama |
  [text854, 3]                       '@AP_Politics | Obama |
  [text968, 6] '@FoxNews@MariaBartiromo Thanks to | Obama |
  [text998, 4]             '@Carydc@gentlemanirish | Obama |
 [text1088, 3]                        '@dcexaminer | Obama |
 [text1098, 2]                          President | Obama |
 [text1099, 2]                          President | Obama |
 [text1112, 2]                          President | Obama |
 [text1281, 9]          @TIME Americans are sick of | Obama |
 [text1288, 1]                                   | Obama |
 [text1289, 1]                                   | Obama |
 [text1359, 6]      '@chicagotribune but that's not | obama |
 [text1484, 2]                          President | Obama |
[text1501, 15]     , hospital funds lawsuit against | Obama |
 [text1560, 4]               '@JohnFromCranber but | Obama |
 [text1573, 2]                 #IFlippedOutBecause | Obama |
[text1584, 13]                     poor! Nice try, | Obama |
 [text1585, 7]            @usacsmret so, looks like | Obama |
 [text1623, 4]                 First lady Michelle | Obama |
 [text1656, 8]    he'd consider assisted#suicide. | Obama |
 [text1716, 3]                          '@mashable | Obama |
 [text1779, 5]              Supreme Court sides with | Obama |
[text1785, 16]          need Obamacare Canny move by | Obama |
 [text1792, 1]                                   | Obama |
```

4. Topic Models, what for

https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf

Topic models extend and build on classical methods in natural language processing such as the unigram model and the mixture of unigram models (Nigam, McCallum, Thrun, and Mitchell 2000) as well as Latent Semantic Analysis (LSA; Deerwester, Dumais, Furnas, Landauer, and Harshman 1990). Topic models differ from the unigram or the mixture of unigram models because they are mixed-membership models (see for example Airoldi, Blei, Fienberg, and Xing 2008). In the unigram model each word is assumed to be drawn from the same term distribution, in the mixture of unigram models a topic is drawn for each document and all words in a document are drawn from the term distribution of the topic. In mixed-membership models documents are not assumed to belong to single topics, but to simultaneously belong to several topics and the topic distributions vary over documents.

An early topic model was proposed by Hofmann (1999) who developed probabilistic LSA. He assumed that the interdependence between words in a document can be explained by the latent topics the document belongs to. Conditional on the topic assignments of the words the word occurrences in a document are independent. The latent Dirichlet allocation (LDA; Blei, Ng, and Jordan 2003b) model is a Bayesian mixture model for discrete data where topics are

4. Topic Models, what for

https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf

assumed to be uncorrelated. The correlated topics model (CTM; Blei and Lafferty 2007) is an extension of the LDA model where correlations between topics are allowed. An introduction to topic models is given in Steyvers and Griffiths (2007) and Blei and Lafferty (2009). Topic models have previously been used for a variety of applications, including ad-hoc information retrieval (Wei and Croft 2006), geographical information retrieval (Li, Wang, Xie, Wang, and Ma 2008) and the analysis of the development of ideas over time in the field of computational linguistics (Hall, Jurafsky, and Manning 2008).

4. Topic Models, what for

2. Topic model specification and estimation

2.1. Model specification

For both models—LDA and CTM—the number of topics k has to be fixed a-priori. The LDA model and the CTM assume the following generative process for a document $w = (w_1, \ldots, w_N)$ of a corpus D containing N words from a vocabulary consisting of V different terms, $w_i \in \{1, \ldots, V\}$ for all $i = 1, \ldots, N$.

4. Topic Models, what for

2.3. Pre-processing

The input data for topic models is a document-term matrix. The rows in this matrix correspond to the documents and the columns to the terms. The entry m_{ij} indicates how often the jth term occurred in the ith document. The number of rows is equal to the size of the corpus and the number of columns to the size of the vocabulary. The data pre-processing step involves selecting a suitable vocabulary, which corresponds to the columns of the document-term matrix. Typically, the vocabulary will not be given a-priori, but determined using the available data. The mapping from the document to the term frequency vector involves tokenizing the document and then processing the tokens for example by converting them to lower-case, removing punctuation characters, removing numbers, stemming, removing stop words and omitting terms with a length below a certain minimum. In addition the final document-term matrix can be reduced by selecting only the terms which occur in a minimum number of documents (see Griffiths and Steyvers 2004, who use a value of 5) or those terms with the highest term-frequency inverse document frequency (tf-idf) scores (Blei and Lafferty

4. Topic Models, what for

2.4. Model selection

For fitting the LDA model or the CTM to a given document-term matrix the number of topics needs to be fixed a-priori. Additionally, estimation using Gibbs sampling requires specification of values for the parameters of the prior distributions. Griffiths and Steyvers (2004) suggest a value of $50/k$ for α and 0.1 for δ. Because the number of topics is in general not known, models with several different numbers of topics are fitted and the optimal number is determined in a data-driven way. Model selection with respect to the number of topics is possible by splitting the data into training and test data sets. The likelihood for the test data is then approximated using the lower bound for VEM estimation. For Gibbs sampling the log-likelihood is given by

$$\log(p(w|z)) = k \log\left(\frac{\Gamma(V\delta)}{\Gamma(\delta)^V}\right) + \sum_{K=1}^{k}\left\{\left[\sum_{j=1}^{V}\log(\Gamma(n_K^{(j)} + \delta))\right] - \log(\Gamma(n_K^{(.)} + V\delta))\right\}.$$

4. Topic Models, how

3. Application: Main functions LDA() and CTM()

The main functions in package **topicmodels** for fitting the LDA and CTM models are LDA() and CTM(), respectively.

```
R> LDA(x, k, method = "VEM", control = NULL, model = NULL, ...)
R> CTM(x, k, method = "VEM", control = NULL, model = NULL, ...)
```

These two functions have the same arguments. x is a suitable document-term matrix with non-negative integer count entries, typically a "DocumentTermMatrix" as obtained from package **tm**. Internally, **topicmodels** uses the simple triplet matrix representation of package **slam** (Hornik, Meyer, and Buchta 2011) (which, similar to the "coordinate list" (COO) sparse matrix format, stores the information about non-zero entries x_{ij} in the form of (i, j, x_{ij}) triplets). x can be any object coercible to such simple triplet matrices (with count entries), in particular objects obtained from readers for commonly employed document-term matrix storage formats. For example the reader read_dtm_Blei_et_al() available in package **tm** allows to read in data provided in the format used for the code by Blei and co-authors. k is an integer (larger than 1) specifying the number of topics. method determines the estimation method used and currently can be either "VEM" or "Gibbs" for LDA() and only "VEM" for CTM(). Users can provide their own fit functions to use a different estimation technique or fit a slightly different model variant and specify them to be called within LDA() and CTM() via the method argument. Argument model allows to provide an already fitted topic model which is used to initialize the estimation.

4. Topic Models, how

Argument `control` can be either specified as a named list or as a suitable S4 object where the class depends on the chosen method. In general a user will provide named lists and coercion to an S4 object will internally be performed. The following arguments are possible for the control for fitting the LDA model with the VEM algorithm. They are set to their default values.

```
R> control_LDA_VEM <-
+    list(estimate.alpha = TRUE, alpha = 50/k, estimate.beta = TRUE,
+         verbose = 0, prefix = tempfile(), save = 0, keep = 0,
+         seed = as.integer(Sys.time()), nstart = 1, best = TRUE,
+         var = list(iter.max = 500, tol = 10^-6),
+         em = list(iter.max = 1000, tol = 10^-4),
+         initialize = "random")
```

The arguments are described in detail below.

4. Topic Models, how

The possible arguments controlling how the LDA model is fitted using Gibbs sampling are given below together with their default values.

```
R> control_LDA_Gibbs <-
+    list(alpha = 50/k, estimate.beta = TRUE,
+         verbose = 0, prefix = tempfile(), save = 0, keep = 0,
+         seed = as.integer(Sys.time()), nstart = 1, best = TRUE,
+         delta = 0.1,
+         iter = 2000, burnin = 0, thin = 2000)
```

`alpha`, `estimate.beta`, `verbose`, `prefix`, `save`, `keep`, `seed` and `nstart` are the same as for estimation with the VEM algorithm. The other parameters are described below in detail.

4. Topic Models, how

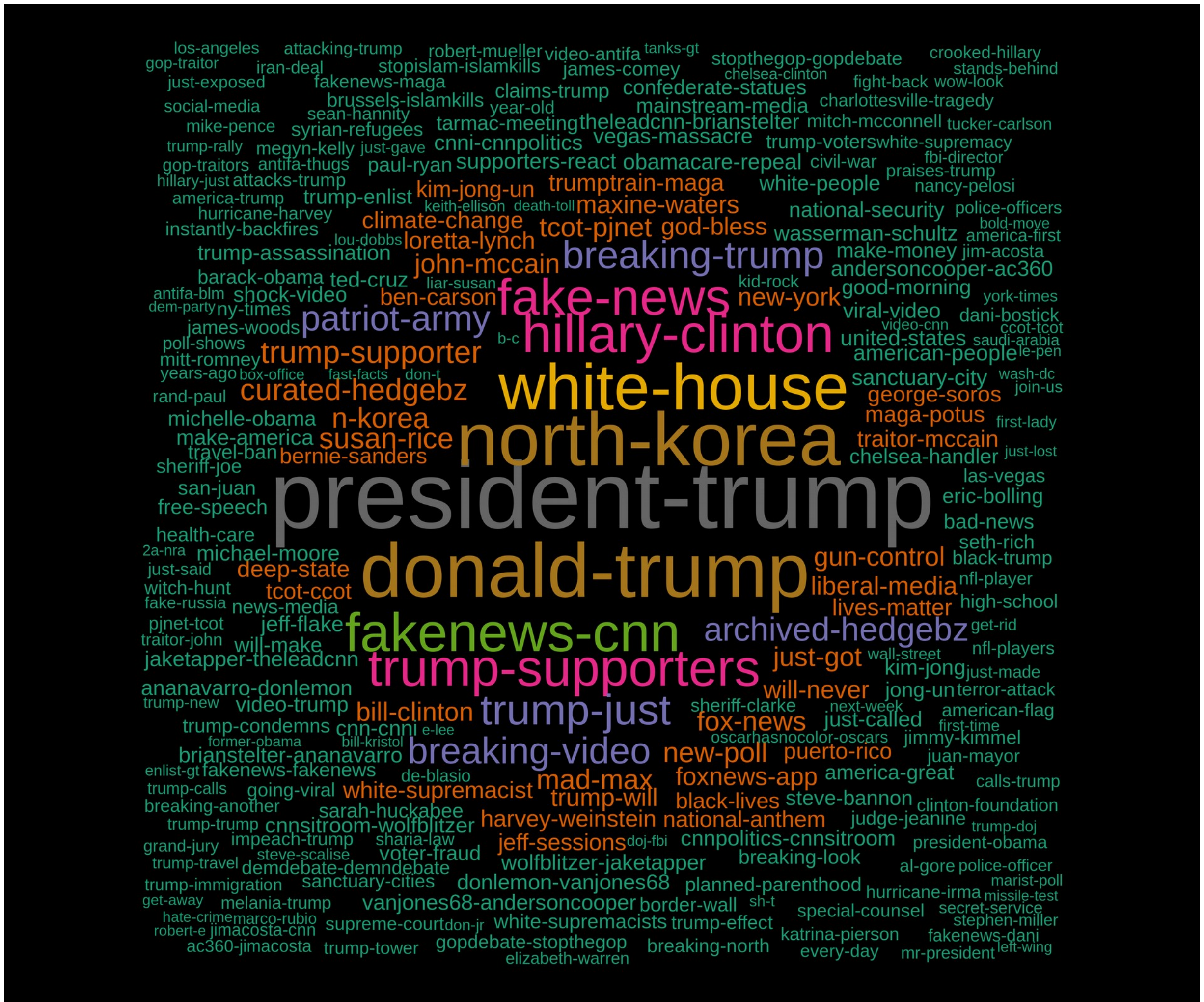
For the CTM model using the VEM algorithm the following arguments can be used to control the estimation.

```
R> control_CTM_VEM <-
+    list(estimate.beta = TRUE,
+         verbose = 0, prefix = tempfile(), save = 0, keep = 0,
+         seed = as.integer(Sys.time()), nstart = 1L, best = TRUE,
+         var = list(iter.max = 500, tol = 10^-6),
+         em = list(iter.max = 1000, tol = 10^-4),
+         initialize = "random",
+         cg = list(iter.max = 500, tol = 10^-5))
```

4. Topic Models, what they return

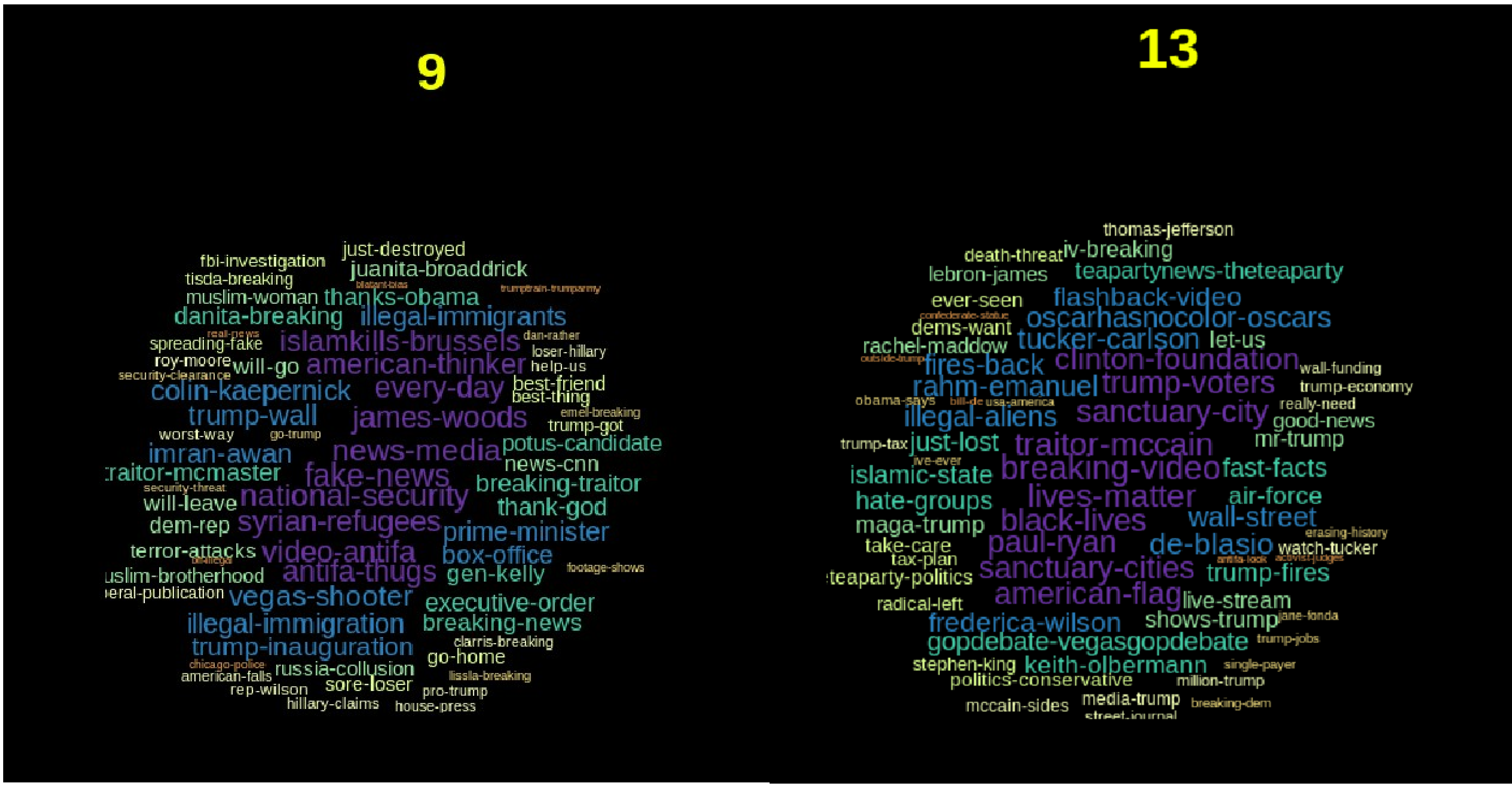
`LDA()` and `CTM()` return `S4` objects of a class which inherits from `"TopicModel"` (or a list of objects inheriting from class `"TopicModel"` if `best=FALSE`). Because of certain differences in the fitted objects there are sub-classes with respect to the model fitted (LDA or CTM) and the estimation method used (VEM or Gibbs sampling). The class `"TopicModel"` contains the call, the dimension of the document-term matrix, the number of words in the document-term matrix, the control object, the number of topics and the terms and document names and the number of iterations made. The estimates for the topic distributions for the documents are included which are the estimates of the corresponding variational parameters for the VEM algorithm and the parameters of the predictive distributions for Gibbs sampling. The term distribution of the topics are also contained which are the ML estimates for the VEM algorithm and the parameters of the predictive distributions for Gibbs sampling. In additional slots the objects contain the assignment of terms to the most likely topic and the log-likelihood which is $\log p(w|\alpha, \beta)$ for LDA with VEM estimation, $\log p(w|z)$ for LDA using Gibbs sampling and $\log p(w|\mu, \Sigma, \beta)$ for CTM with VEM estimation. For VEM estimation the log-likelihood is returned separately for each document. If a positive `keep` control argument was given, the log-likelihood values of every `keep` iteration is contained. The extending class `"LDA"` has an additional slot for α, `"CTM"` additional slots for μ and Σ. `"LDA_Gibbs"` which extends class `"LDA"` has a slot for δ and `"CTM_VEM"` which extends `"CTM"` has an additional slot for ν^2.

4. Topic Models, example right trolls, input bigrams

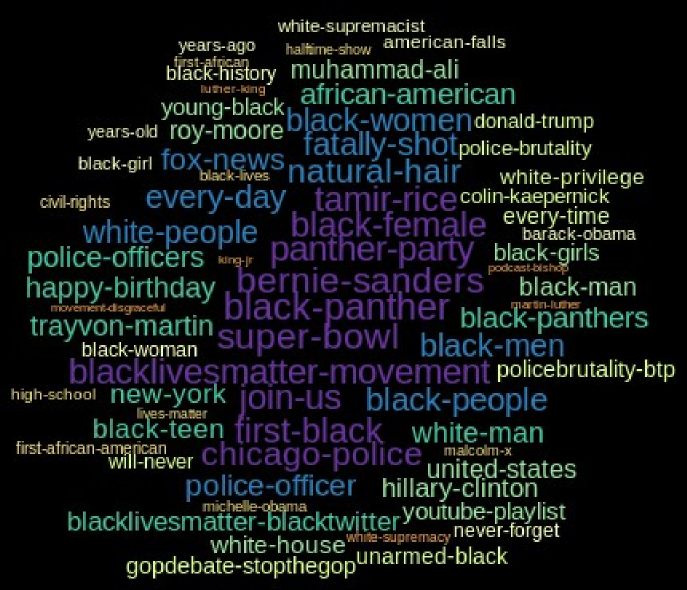


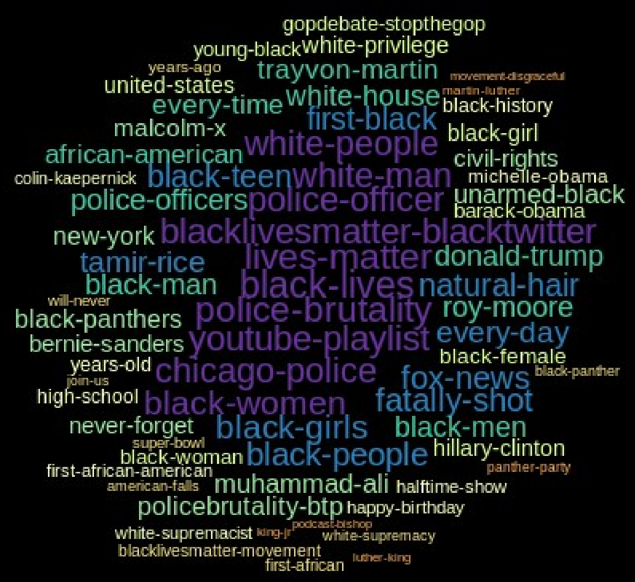
4. Topic Models, example LEFT trolls, input bigrams

4. Topic Models, 2 example topics (of 20) right trolls



4. Topic Models, 2 example topics (of 10) left trolls





References

Text Mining with R- A Tidy Approach https://www.tidytextmining.com/

https://bookauthority.org/books/best-text-mining-books

https://en.wikibooks.org/wiki/R_Programming/Text_Processing

https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

https://quanteda.io/

https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf

https://rstudio-pubs-static.s3.amazonaws.com/266565_171416f6c4be464fb11f7d8200c0b8f7.html

https://cran.r-project.org/web/packages/SentimentAnalysis/vignettes/SentimentAnalysis.html

¡Gracias!

Pedro Concejero
pedro.concejero@u-tad.com
pedro.concejerocerezo@gmail.com
twitter: https://twitter.com/concejeropedro