CRWN 88

Homework 3

Who is Satoshi?

Anders Poirel

October 27, 2018

1 Methods

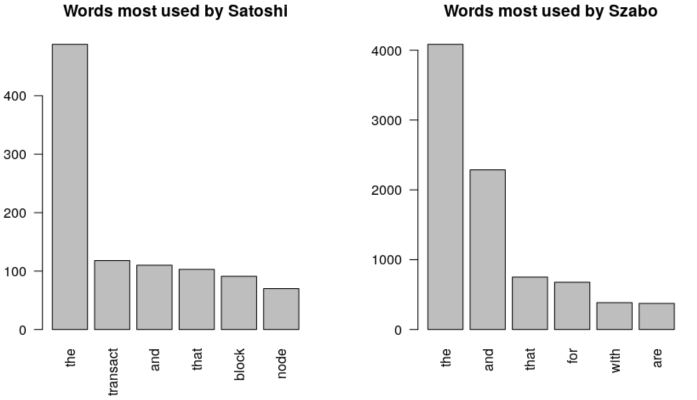
To examine whether the pages from Nick Szabo's blog *Unenumerated* support the hypothesis that he and Satoshi Nakamoto are the same person, we use three different text analysis techniques. In all cases, the two corpuses of documents are on one hand Satoshi's original white paper and twelve of his emails, and on the other a selections of 42 posts on Szabo's blog.

Term Frequency The first two techniques used are variants of the "bag of words" approach, that is, we examine only the frequencies of the terms in each corpus, disregarding grammar or larger patterns of words. We begin by examining the term frequencies (*tf*) in each corpus, and compare the terms most frequently used by each author. Text pre-processing includes stemming of words (that is, reduccing each word to its radical) and removal of stopwords (common connectives such as "and" or "that"). We perform the analysis both with and without the latter, as each case may reveal different aspects of each author's style.
Next, we examine the most frequent words againg, but using the "Term frequency - Inverse document frequency"(*tf-idf*) term-weighting scheme, which adjusts for the fact that some words appear more frequently in general, by offsetting the frequency of each term by the number of documents in the corpus that contain the term.
In both cases, the similarity (or lack thereof) of the lists of most frequent terms used by each author will be used to evaluate the hypothesis that both authors are the same person.

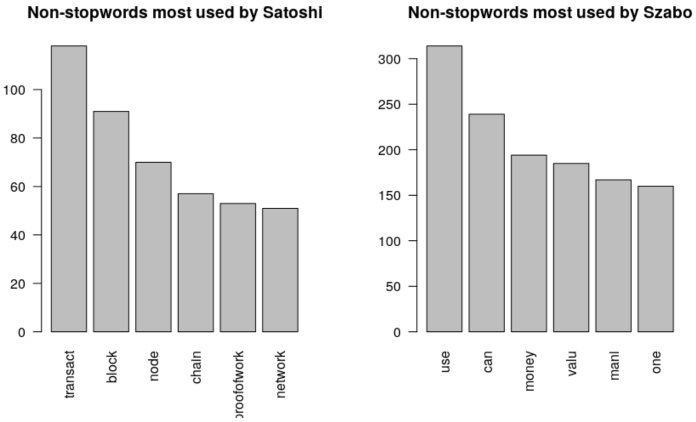
Hierarchical clustering Finally, using a corpus containign both texts from Satoshi and Sazbo, texts are grouped by similarity into clusters. The clustering algorithm used ("Ward's method") iteratively clusters the texts by relative distance (*distance* here is a measure of the relative difference between two texts), merging the texts that are closest to each other for each distance. For a more detailed explanation, see here.

2 Results

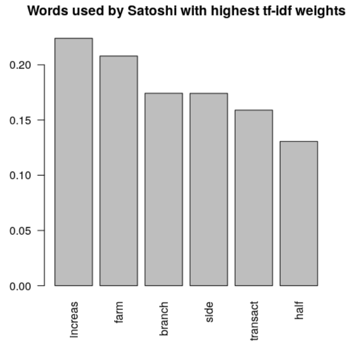
Term frequency without stopword removal:

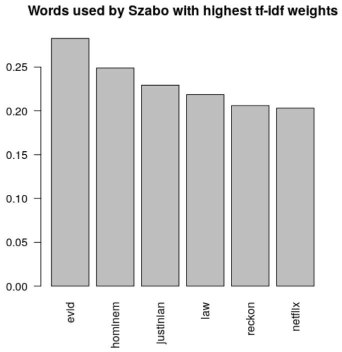


Term frequency with stopword removal:

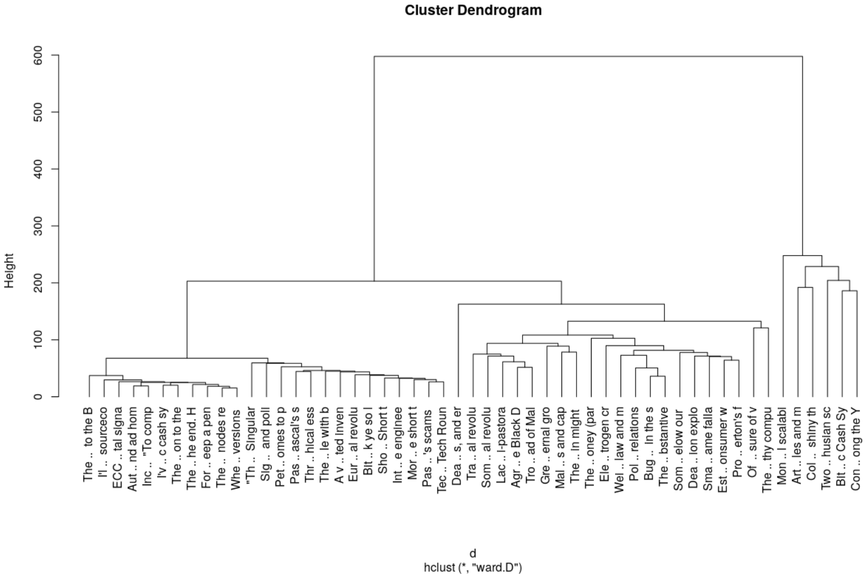


Term-frequency - Inverse document frequency:





Hierarchical cluster analysis:



3 Discussion

Interpretation When stopwords are included term-frequency analysis suggests differences in Satoshi's and Szabo's writing styles. While only three of the most frequent terms for Satoshi are stopword ("the", "and" and "that"), all six of Szabo's most frequently used terms are in this category. This imply that Satoshi tends to write shorter, more concises sentences while Szabo's are longer and wordier, requiring the use of many connectives.
Once stopwords are removed, we see that Satoshi frequently uses precise technical terms - "block", "chain" and "proof of word", while Szabo often speaks of more abstract concepts - "money" and "value". However this is not so much indicative of difference in authoraship as in topics. The texts in the Satoshi sample are focused on the technical implementation of bitcoin, whereas Szabo's blog frequently discusses the interactions between money and society.
Simalarly, the *tf-idf* results shows technical terms pertaining to blockchain for Satoshi - "branch", "transaction" and "farm", and on Szabo's side mostly temrs pertaining to argumentation and politics - "evidence", "hominem", "law". Again, this highlights that the texts in each sample treat different topics, but does not offer any evidence as to whether they have the same author.

Finally, Satoshi's texts are almost all clustered on the far left of the dendogram, with a small relative distance bewteen each other (< 50) and a larger relative difference with most of Szabo's texts (200). This suggests that there is in general little ressemblance between his texts and Szabo's in the sample. On the flipside, depsite having selected the longest possible samples of Satoshi's writing, all were considerably shorter and less "wordy" than Szabo's often in-depth blog posts, so it seems natural that the algorithm would find little similarity between both corpuses. A notable outlier to this was Satoshi's white paper, "Bitcoin: A Peer-to-Peer Electronic Cash System", which is clustered with Nick Szabo's blog posts about different monetary systems, ecomomic value and other similar topics. While this might be used to suggest that Nick Szabo is the author of that paper, the larger relative distance (≤ 200) between the paper and other items in the cluster make this weak evidence a best. Furthermore, it is likely that the algorithm clustered these items together due common keywords such as "exchange", "value" or "money", which indicates common topics but not authorship.
Overall, these results thus offer little evidence that Satoshi Nakamoto and Nick Szabo are in fact the same person.

Limitations For practical purposes, not all texts from the Satoshi Archive nor Unenumerated were used, given that in both cases, hundreds of pages were available, which would have been impractical to process by hand[1]. Those that were used were selected rather arbitrarily. For Szabo, all blog posts predating 2010 were ignored, as were his responses to comments posted on his blog. For

[1] I did look into implementing a web scaper in Python then R, but I essentially ran out of time to do so

Satashi, only his original white paper and a selection his longer emails were considered, a choice motivated by the fact thar his forum posts were often very short, lending themselves poorly to comparison with Szabo's long blog posts. Thus, the data used was neither complete nor sampled in a representative manner, limiting the validity of the results. Also, only basic text analysis techniques were used, as I lacked the knowledge to implement more sophisticated (and informative?) techniques. For instance, examining patterns of two words or more at a time could have reveal more about the similarities betweeen Satoshi's and Szabo's writings.

4 R implemementation

```
#------------------------------------------------------------------------
# dtm.r
# finds words most and least frequently used by both Satoshi and Szabo in
# a sample of texts from both authors
# then does the same using a tf-idf weighting
#
# Running instructions:
#   > setwd(PATH) where PATH is the path of the directory containing
    dtm.R
#             and the three text folders
#   > source(dtm.R)
#------------------------------------------------------------------------

library(tm)
library(ggplot2)

# function for pre-processing of corpuses
clean_corpus <- function(corpus, remove_stop) {
    corpus <- tm_map(corpus, removePunctuation)
    corpus <- tm_map(corpus, toSpace, "")
    corpus <- tm_map(corpus, toSpace, "")
    corpus <- tm_map(corpus, toSpace, " -")
    corpus <- tm_map(corpus, toSpace, "")
    corpus <- tm_map(corpus, toSpace, "")
    corpus <- tm_map(corpus, content_transformer(tolower))
    corpus <- tm_map(corpus, removeNumbers)
    corpus <- tm_map(corpus, stripWhitespace)
    if(remove_stop) {
        corpus <- tm_map(corpus, removeWords, stopwords("english"))
    }
    corpus <- tm_map(corpus, stemDocument)

    return(corpus)
}
```

```
#creates corpus of Satoshi's and Szabo's writings
satoshi <-
    Corpus(DirSource("/home/joachim/Documents/SchoolWork/CRWN88/Satoshi"))
szabo <-
    Corpus(DirSource("/home/joachim/Documents/SchoolWork/CRWN88/Szabo"))

# pre-processing
toSpace <- content_transformer(function(x, pattern) {return
    (gsub(pattern, "", x))})
a <-readline("Remove stopwords for frequency analysis? (y/n) :")
if (a == "yes" | a == "y") {
   remove_stop <- TRUE
} else {
   remove_stop <- FALSE
}
satoshi <- clean_corpus(satoshi, remove_stop)
szabo <- clean_corpus(szabo, remove_stop)


# -------------------------
# Simple frequency analysis
# -------------------------

# builds and orders Document term matrices
dtm_sza <- DocumentTermMatrix(szabo)
dtm_sat <- DocumentTermMatrix(satoshi)
freq_sat <- colSums(as.matrix(dtm_sat))
freq_sza <- colSums(as.matrix(dtm_sza))
ord_sza <- order(freq_sza, decreasing = TRUE)
ord_sat <- order(freq_sat, decreasing = TRUE)


# display 6 most and least frequently used words for satoshi and szabo
print("Words most used by Satoshi:")
print(freq_sat[head(ord_sat)])
print("Words least used by Satoshi:")
print(freq_sat[tail(ord_sat)])
print("Words most used by Szabo:")
print(freq_sza[head(ord_sza)])
print("Words least used by Szabo:")
print(freq_sza[tail(ord_sza)])

# print histograms of 6 most frequent words used by each author
barplot(freq_sza[head(ord_sza)], main = "Words most used by Szabo", las
    = 2)
barplot(freq_sat[head(ord_sat)], main = "Words most used by Satoshi",
    las =2)
```

```
# ---------------
# Tf-idf analysis
# ---------------

tfidf_dtm_sat <- TermDocumentMatrix(satoshi, control = list(weighting =
    weightTfIdf))
tfidf_dtm_sza <- TermDocumentMatrix(szabo, control = list(weighting =
    weightTfIdf))
tfidf_freq_sat <- rowSums(as.matrix(tfidf_dtm_sat))
tfidf_freq_sza <- rowSums(as.matrix(tfidf_dtm_sza))
tfidf_ord_sat <- order(tfidf_freq_sat, decreasing = TRUE)
tfidf_ord_sza <- order(tfidf_freq_sza, decreasing = TRUE)

# display 6 word with highest tf-idf weight for each author
print("Words used by Satoshi with highest tf-idf weight:")
print(tfidf_freq_sat[head(tfidf_ord_sat)])
print("Words used by Szabo with highest tf-idf weight:")
print(tfidf_freq_sza[head(tfidf_ord_sza)])

# print histograms of 6 words with highest tf-idf weight for each author
barplot(tfidf_freq_sat[head(tfidf_ord_sat)], main = "Words used by
    Satoshi with highest tf-idf weights", las = 2)
barplot(tfidf_freq_sza[head(tfidf_ord_sza)], main = "Words used by Szabo
    with highest tf-idf weights", las = 2)


# --------------------------------
# Hierarchical clustering analysis
# --------------------------------

# pre-processing
alltexts <-
    Corpus(DirSource("/home/joachim/Documents/SchoolWork/CRWN88/Alltexts"))

a <-readline("Remove stopwords for cluster analysis? (y/n) :")
if (a == "yes" | a == "y") {
   remove_stop <- TRUE
} else {
   remove_stop <- FALSE
}
alltexts <- clean_corpus(alltexts, remove_stop)

# builds dtm and prepares for clustering
dtm_all <- DocumentTermMatrix(alltexts)
m_all <- as.matrix(dtm_all)
write.csv(m_all, file = "dtmEight2Late.csv")
rownames(m_all) <-
    paste(substring(rownames(m_all),1,3),rep("..",nrow(m_all)),
    substring(rownames(m_all), nchar(rownames(m_all))-12,
    nchar(rownames(m_all))-4))
```

```
#compute distance between document vectors
d <- dist(m_all)
#run hierarchical clustering using Ward's algorithm
groups <- hclust(d,method="ward.D")

# plot dendogram
plot(groups, hang = -1)
```

5 Acknowledgements

Sources of texts authored by Satoshi and Szabo:

- “The Complete Satoshi”
  https://satoshi.nakamotoinstitute.org/
- “Unenumerated”,
  https://unenumerated.blogspot.com/

Methods used for text mining in R:

- “A gentle introduction to text mining using R”,
  https://eight2late.wordpress.com/2015/05/27/a-gentle-introduction-to-text-mining-using-r/
- “Text Mining: Bag of Words”,
  https://rpubs.com/williamsurles/316682