University of Glasgow | School of Computing Science

Tracing malicious behaviour on Twitter

George Illingworth

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Level 4 Project — March 28, 2018

Abstract

I construct a series of algorithms that compare the semantics, syntax and grammar of Twitter users. These algorithms are then tested on their ability to distinguish the author of the Tweets.

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: ____________________ Signature: ______________________

Contents

1 Introduction **1**

1.1 Background 1

1.2 Motivation 2

2 Taking A Linguistic Fingerprint **4**

2.1 Forming the Fingerprint: Theory 4

2.1.1 Naive Bag-of-Words Approach 4

2.1.2 Part-of-Speech Approach 4

2.1.3 Tree-building approach. 6

2.2 Forming a Fingerprint: Technical Detail 8

2.2.1 Data Collection 8

Chapter 1

Introduction

1.1 Background

Next to search engines, social networks form the backbone of the modern internet. The SimilarWeb records[**?**] for the most used websites worldwide place various social networks in 7 of the top 50 places. This list places 24 search engines in the top 50, but the overwhelming majority of these (21 of the 24) are various international versions of Google. As the usage of social networks increases across the globe, so too increases their societal and cultural power. Social networks represent a fundamental change in the way news, entertainment and other media are processed, produced and shared.

At one time there was a small subset of the population that was responsible for the vast majority of content creation. Monolithic institutions like newspapers, radio stations or television networks provided the news of the day in an edited, concise and condensed form. These institutions have always made very deliberate decisions about the sort of content that they choose to provide their audience; every article in every newspaper was, and is, run through multiple layers of editing to ensure the tone, style and content are consistent with the image that they want to present. To become a writer, to be allowed to produce material for the general population to consume, required years of training, education and work experience.

The explosion of social networks changed all of this. As a user of a social network, everyone, broadly speaking, has exactly the same rights to produce the content they want as everyone else. This freedom applies to every facet of the content creation process. It is possible to imagine a writer for a newspaper having carte blanche to write about any subject they see fit, but they will still be beholden to certain style guidelines. They would not, for example, be allowed to publish full articles using only capital letters, or to misuse and mangle punctuation and spelling for narrative effect.

In addition, every user can decide exactly how much of any other user's content they want to consume at any given moment. If one user likes what another is doing, they are free to copy or share it as they see fit. This allows communities to be built and messages, ideas and sentiments to be spread all over the world at unimaginable speeds.

These ideas may present little cultural value beyond temporary amusement - it is hard to argue that social phenomena such as the 'planking' fad of 2011 have left any lasting mark on society - but they can also be very influential in enacting huge changes in national and international politics.

In the beginning of the same year, Hosni Mubarak was ousted as President of Egypt after holding the position for nearly 30 years.[1] Social networking platforms in this case gave voices to those who would otherwise not have been afforded them, and enabled them a much greater reach than traditional word-of-mouth organisation.

The same tools, generously provided by social networks, have been used to similar effect in Ukraine[2], Iran[3] and Tunisia[4]. While it would be overly simplistic to state that none of these events would have happened without social networks, their influence was strongly and incontrovertibly felt.

There is significant value and power in the communities and the discourse which are made possible through these networks, and where there is power there are going to be people who want to control it. In some cases this may be purely selfish; through the creation of several sock puppet accounts, it can be possible for a user to construct notoriety and fame for themselves. Possibly the best documented of these phenomena occurred within a large community of Harry Potter fans, spread across many websites between 2003 and 2006. Fiction Alley[1] user 'msscribe'[5] created and maintained a series of profiles across many similar sites that would both praise and deride her work. Through the intentional creation of controversy, she managed to achieve significant status within that community.

This sort of behaviour has continued in different, and in many ways more sinister, forms as the global influence of social networks has grown. Rather than individuals simply trying to create a name for themselves, political groups and even nation states have been shown to use similar tactics in attempts to exert influence in their own countries and abroad. This is now an expected part of the news cycle for any major political event that has occurred in the last few years. The United States Presidential Election of 2016[6] , Brexit[7] and the French Presidential Election of 2017[8] all featured some form of attempted tampering from outside sources via groups of directed social networking accounts.

There is definite tangible value in being able to detect when this sort of attack is happening in an automated manner, but this can often be very complex. The difficulty involved in detecting such groups of accounts depends heavily on how they are created and how they are being run. In simple cases where no effort is made to disguise the location of the individual or group who is orchestrating the attack, they can be traced via their IP address, but it is trivial for these addresses to be faked or hidden. Other factors that can be used to trace such activity include the creation of many similar accounts in a short time period, often rudimentary approaches to grammar, repetitiveness and formations of unusual posting patterns between accounts.[6] All of these hallmarks can be obscured in some form, and most are negated if each account is being controlled manually.

1.2 Motivation

The democratization of content creation allows every user to form their own style, where they can choose to use - or ignore - any of the customs of any language they want. While in traditional media a writer is encouraged to find their own voice within strict grammatical guidelines, the form and shape of the content produced by a social media user is allowed to be considerably more individualized.

In traditional writing, there is posited to exist a sort of *linguistic fingerprint*[**?**], where each writer has their own style that can be positively identified. The applicability of this rule relies heavily on the context: successfully identifying a single writer from a single piece of text is highly unlikely and borders on the absurd. It is relatively well established that a writer's *fingerprint* changes over time[**?**],

This paper examines the hypothesis that each social media user leaves behind some kind of *linguistic fingerprint*; and that, through focusing purely on the textual content of an account's posts, it is possible to determine the author. Moreover, a series of algorithms designed to extract such a fingerprint are designed, tested and analyzed.

These algorithms were designed around Twitter users, for a number of reasons. Twitter is the third most popular social media site worldwide[**?**]; it grants considerably more access through APIs to outside developers; and it sets a strict, small character limit per post. This limit - historically 140 characters but in recent times

[1] A now archived website for users to post their own writings that exist within and expand upon the Harry Potter universe.

moving to 280 - both encourages unusual linguistic behaviour, as users try to encapsulate as complete an idea as possible within the limit, and provides relatively small blocks, which can be analysed independently. The length constraints on each post allow them to be considered individually in a more performance-efficient and meaningful manner. Two of the three algorithms proposed run in at least linear time with respect to the length of the average post, and so a low upper-bound on that length keeps the run times reasonable.

As discussed, there can be no expectation that a user will punctuate correctly, and so attempting to distinguish sentences automatically can in many cases present an extremely complex task. However, the character limit imposed by Twitter means that users are forced to split long and complex ideas they may wish to write about across many posts. This means most individual posts contain only a few key ideas, and so there is a reasonable equivalence to be made between a single Twitter post and a single sentence. This is extremely helpful as the deconstruction of a sentence to its constituent grammatical parts is a common linguistic task.[9]

Chapter 2

Taking A Linguistic Fingerprint

2.1 Forming the Fingerprint: Theory

2.1.1 Naive Bag-of-Words Approach

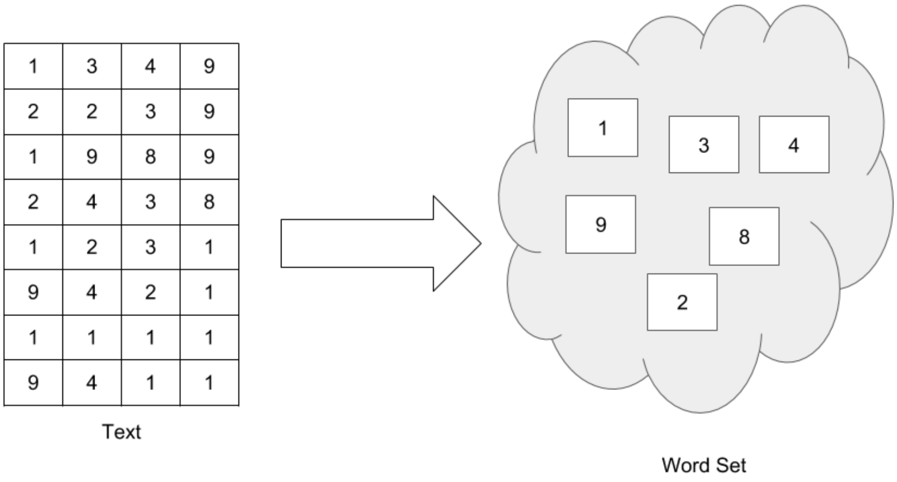
The simplest of the methods chosen to determine a fingerprint for a given user starts by forming a set of all the words present in the user's posts. In this method, the term 'Word' simply denotes any sequence of characters separated from those around it by spaces. This could include URLs, mentions of other usernames, emoticons, mis-spellings and a variety of other non-dictionary words, but these were not filtered, as they also contribute to the fingerprint of an individual. Since a set is a group of unique elements with no defined order, this set formation turns a list of posts - which can be considered as lists of words - into a disordered group of words. A representation of this process is shown in 2.1, where the lists of words have been replaced with a table of numbers for ease of reading.

These sets can then be compared to each other by examining the ratio between the sizes of the original sets (A and B in 2.2) to the size of the common set (C). If this ratio is unusually high, it is hypothesized that the author of each set is likely to be the same person.

2.1.2 Part-of-Speech Approach

The Bag-Of-Words approach detailed above focuses only on the word choice of a given account, but the grammatical structures may be just as revealing. Sentences are made up of several constituent parts: they should contain at least a noun and a verb, but almost all contain more words with different, more complex uses. Of course, Twitter users are not constrained to writing in complete sentences or even to attempting to impart meaning in this way. In practice, however, it is found that the vast majority of users are attempting in some way to transmit some form of information, but this is often done in ways that fail to comply with, or even actively disregard, traditional ideas of grammar and structure.

This could potentially cause issues when attempting to decipher the structures being used, as the vast majority of tools for deciphering this sort of meaning are designed mainly with various formal writings in mind. That being said, when writers are not constrained by the expectation of using strict grammar and convention, they are likely to create or adopt their own. These new conventions are likely to enhance the fingerprint and are detectable with the right tools. This is especially true when considering the space constraints enforced by Twitter: if a post is slightly too long, the user is likely to look through what they have written for any superfluous words they can remove entirely without the post losing meaning. By doing this, the user is forced to make their own decisions



1	3	4	9
2	2	3	9
1	9	8	9
2	4	3	8
1	2	3	1
9	4	2	1
1	1	1	1
9	4	1	1

Figure 2.1: Graphic representation of set formation.

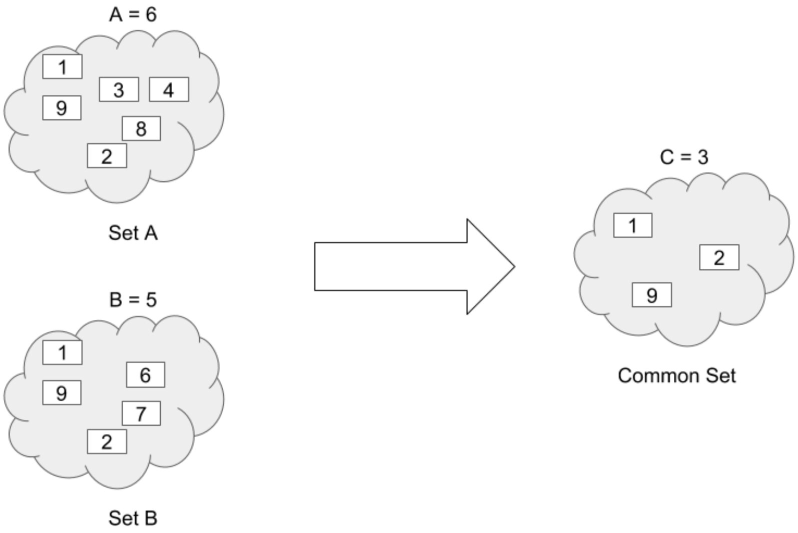
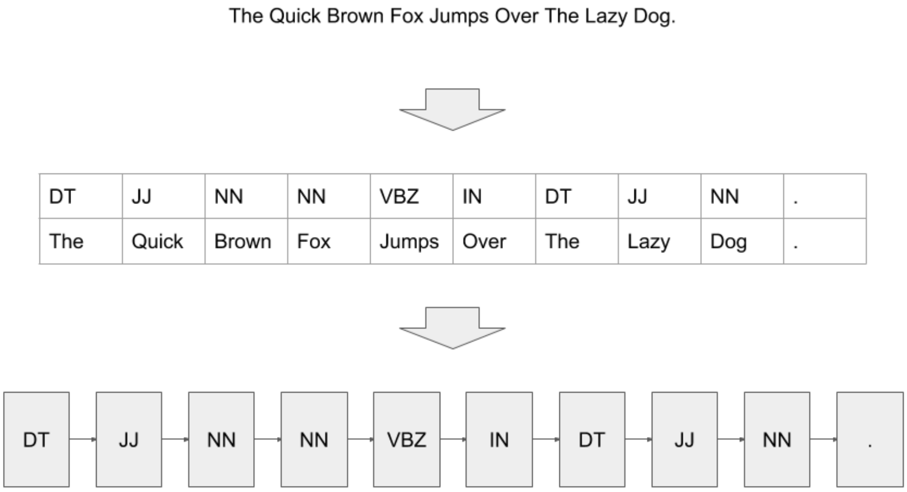


Figure 2.2: Graphic representation of set comparison.



The Quick Brown Fox Jumps Over The Lazy Dog.

DT	JJ	NN	NN	VBZ	IN	DT	JJ	NN	.
The	Quick	Brown	Fox	Jumps	Over	The	Lazy	Dog	.

Figure 2.3: Graphic representation of tagging.

about what they consider to be important in a sentence, and these choices are likely to be relatively individual. When a user is forced to make these sorts of decisions regularly, patterns should begin to emerge and these should be detectable.

There is not a simple 1:1 relationship between a word and its meaning in a syntactic sense. For example, consider the sentence ”I refuse to collect the refuse.” The word ”refuse” is used both as a noun and a verb, so the context is important in deciphering which meaning is most appropriate at what time. Deciphering meaning in this sense can be computationally complex, but as discussed in **??**, there are freely available tools that can perform most of the heavy lifting for us. This process is called tagging and the tools that do it are known as taggers.

Tagging a post leaves us with a list of the parts of speech used within it, and these lists can be compared to one another in a number of ways. The metrics that were hypothesised to provide greatest insight were the edit distance (ED) and the longest common subsequence (LCS). If posts are broadly similar in makeup, then this is tracked by the ED, and if they reuse certain smaller structures, these are tracked by the LCS. This provides us with effective tools for comparing one post to another, but these measures are statistical and noisy and therefore only of any real use in aggregate.

The chosen method of aggregation when comparing two sets of posts is to compare the means of these measurements within and between the sets. 2.6 illustrates the comparisons required to evaluate two sets with three elements each. If we are to compare the sets of posts A and B, the first step is to find the mean ED and mean LCS between one post in A and another also from A. Next the mean ED and mean LCS between posts from B are taken, before finally the same measurements are performed between the sets.

The mean length of each list in each set was also recorded, as longer lists tend to increase both ED and LCS. It is theorized that the mean ED and mean LCS should be lower in the within-set cases than the between-set case when they have different authors, and that if they are very similar then they have the same author.

2.1.3 Tree-building approach.

As a fundamental function of written language, the words in a post are arranged in a linear fashion. There is value in understanding language in this way, but the underlying structures cannot be fully expressed in one dimension.

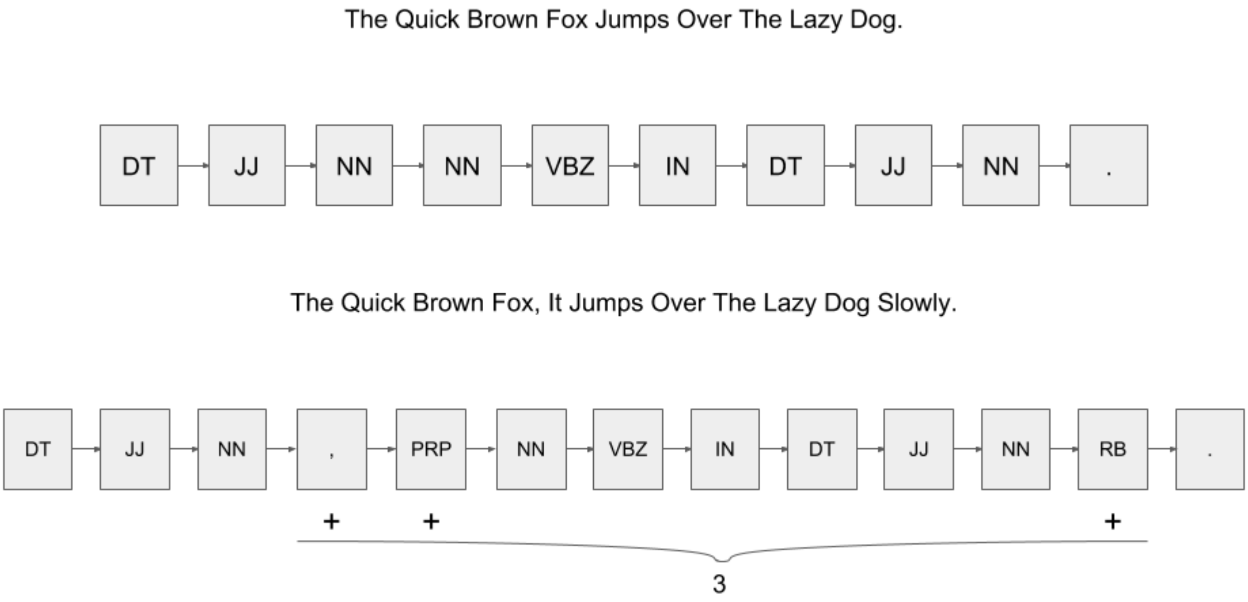

The Quick Brown Fox Jumps Over The Lazy Dog.
DT
JJ
NN
NN
VBZ
IN
DT
JJ
NN
.
The Quick Brown Fox, It Jumps Over The Lazy Dog Slowly.
DT
JJ
NN
,
PRP
NN
VBZ
IN
DT
JJ
NN
RB
.
+
+
+
3

Figure 2.4: Edit Distance Capture.

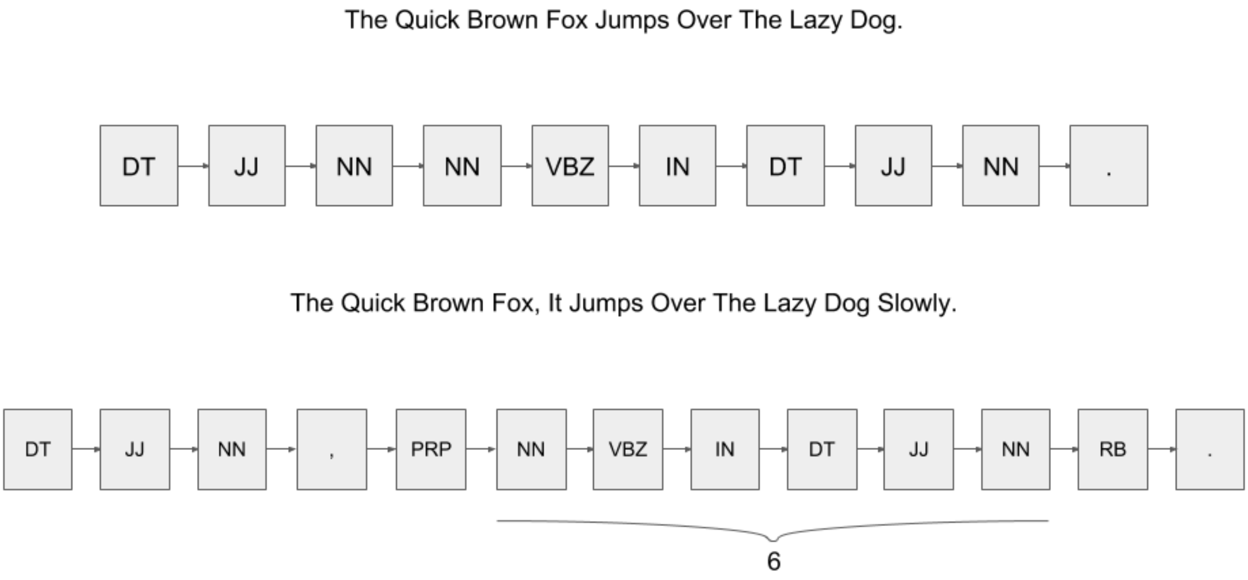

The Quick Brown Fox Jumps Over The Lazy Dog.
DT
JJ
NN
NN
VBZ
IN
DT
JJ
NN
.
The Quick Brown Fox, It Jumps Over The Lazy Dog Slowly.
DT
JJ
NN
,
PRP
NN
VBZ
IN
DT
JJ
NN
RB
.
6

Figure 2.5: Longest Common Subsequence Capture.

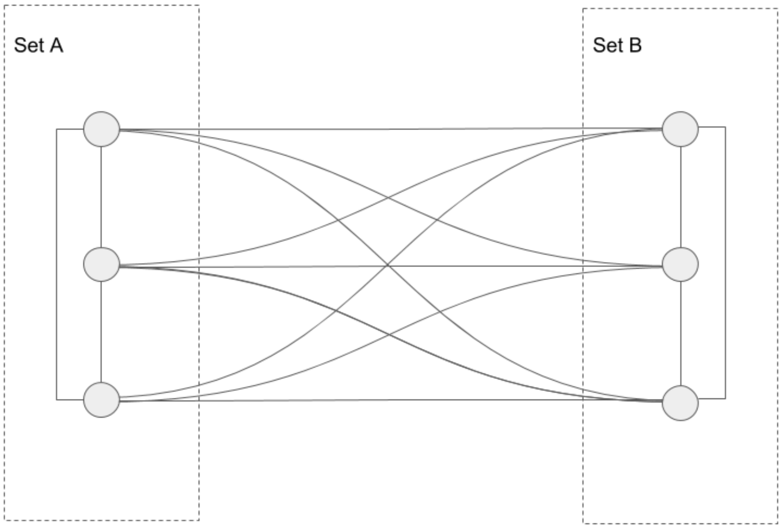


Figure 2.6: Comparisons performed within and across 2 sets.

Writing, as it is used and understood, does not function like a list; instead, words are grouped with common parents, and these groups are combined themselves and so on and so forth into a single entity. This constitutes a bottom-top explanation of a tree, and at the sentence level this is how language is most accurately represented.

Representing the structure of a sentence as a list of parts can be valuable and is in many senses accurate, but it is not a complete picture. A word relates to more than just the words before and after it: they form groups within themselves as phrases or clauses, and these groups can themselves form larger structures which eventually build into a complete sentence. This can be represented as a tree (as shown in 2.7 and 2.8), where each of the words is a leaf and every grouping is a parent node.

There are freely available tools which can parse the sentence into these tree structures; and although it is more computationally expensive than for lists, trees can compared by edit distance. It is also possible to calculate the largest common subtree - which is analogous to the longest common subsequence used in 2.1.2 - but due to the increased complexity and therefore time required for ED in this form, the LCS was not calculated. This means that the methods for obtaining similarity can follow the same patterns laid out in 2.1.2. When given two sets of posts, a tree is generated for each post in each set, and the ED between it and every other tree is calculated. The mean ED values within and between sets are recorded. As in 2.1.2, it is hypothesized that if the author of each set is the different then there should be a higher mean ED between sets, and if they have the same author there should be no significant difference.

2.2 Forming a Fingerprint: Technical Detail

2.2.1 Data Collection

In order to build a system which can test the hypotheses, it is firstly necessary to build an appropriate data set. In this case the data set should consist of a series of post

*****THIS IS AS FAR AS I HAVE WRITTEN****

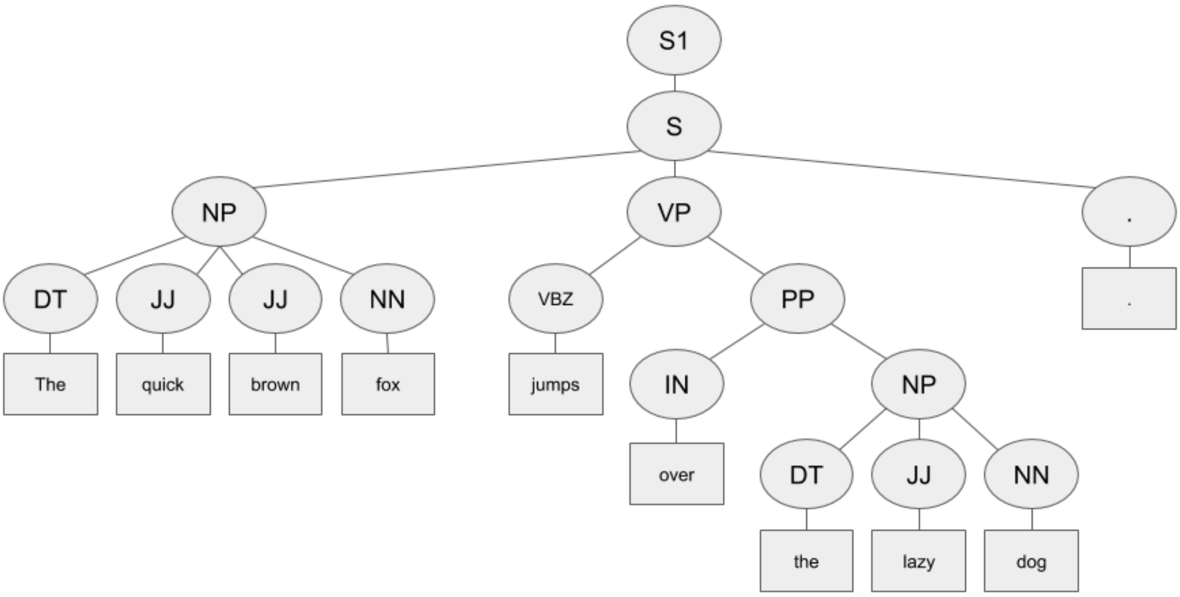


Figure 2.7: Tree formation for ”The quick brown fox jumps over the lazy dog”.

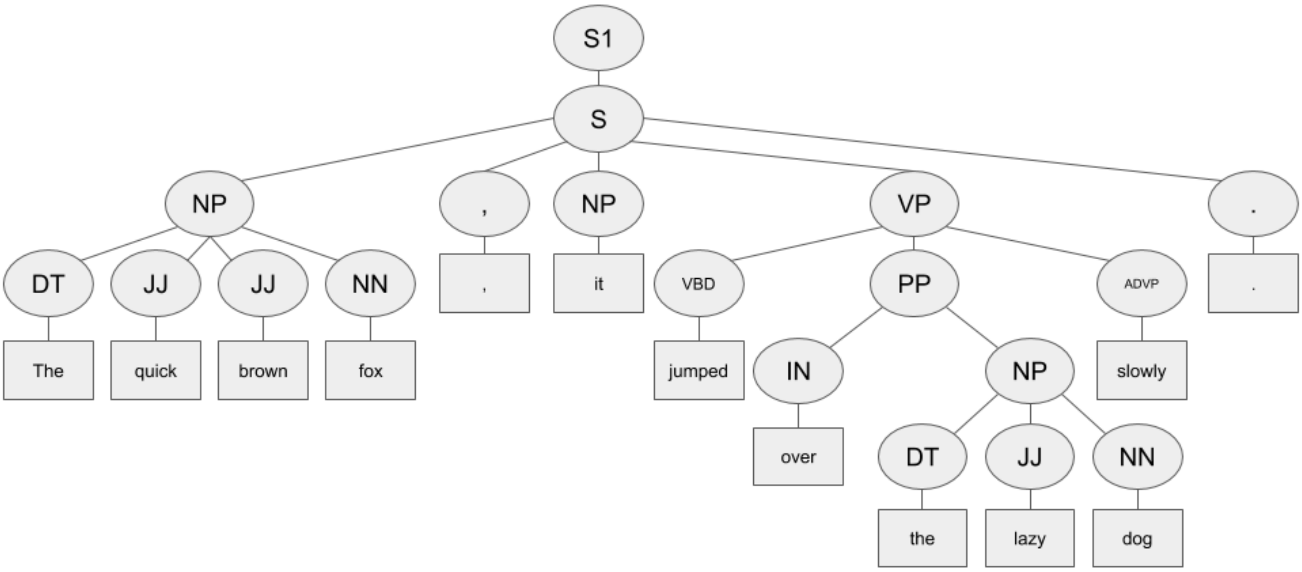


Figure 2.8: Tree formation for ”The quick brown fox, it jumped over the lazy dog slowly”.

Bibliography

[1] Maeve Shearlaw.
Egypt five years on: was it ever a 'social media revolution'?.
The Guardian, 25th February 2011.

[2] Jennifer Dickinson.
Prosymo Maksymal'nyi Perepost! Tactical and Discursive Uses of Social Media in Ukraine's Euromaidan.
Ab Imperio, 3/2014:75–93, 2014.

[3] Lev Grossman.
Iran Protests: Twitter, the Medium of the Movement.
Time, 17th June 2009.

[4] Peter Beaumont.
The truth about Twitter, Facebook and the uprisings in the Arab world.
The Guardian, 25th February 2011.

[5] charlottelennox (pseudonym)
The Ms.Scribe Story: An Unauthorized Fandom Biography.
Journalfen (archived), February 2007.

[6] Scott Shane.
The Fake Americans Russia Created to Influence the Election.
New York Times, 7th September 2017.

[7] Robert Booth, Matthew Weaver, Alex Hern , Stacee Smith and Shaun Walker.
Russia used hundreds of fake accounts to tweet about Brexit, data shows.
The Guardian, 14th November 2017.

[8] Emilio Ferrara.
Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election.
First Monday, 22(8), 2017

[9] Amy Reynolds.
Drawing Sentence Syntax Trees.
Ling 101 Online,
`http://amyrey.web.unc.edu/classes/ling-101-online/tutorials/how-to-draw-syntax-trees/`