Big Data Analysis

Project

Yoga tweets and articles on the internet

Olga Valls Murcia

Introduction

This project has been structured in two parts in order to analyse how yoga is seen on the internet.

A couple of the main questions that could be answered would be:

- Is yoga the soft version of pilates?
- Is it viewed as something more than just stretching the muscles?

The main parts of the project are:

- Part 1: data gathering and storage from #yoga tweets. The idea in this part is to analyse the languages of the tweets for this topic, how the content is generated, the most common secondary hashtags and the most common words (nouns) used in the texts of the tweets.

- Part 2: Yoga articles from yogamag.net and healthandyoga.com websites. The idea for this part is to analyse which are the most popular articles for each of the websites as well as the most common words (nouns) used in the titles of those articles.

After presenting the analysis for each of the parts, the conclusion and the information about the git repository can be found.

PART 1: Data gathering and storage from #yoga tweets

For the first part of this practical work, the #yoga hashtag has been used, and a total number of **12362 tweets** have been captured and stored in a Mongo Database.

From these tweets, the plots shown in this section have been generated, as well as recycled from the code provided in the BigDataTwitterLab.pdf.

Languages spoken in the tweets captured

English is the most common language for the tweets (8637), being almost 70% of the total. Following, with very low percentage, are spanish (661 tweets) with the 5,32% of the total and japanese, russian, hindi, french, portugueese with less than 3% of the total.

More than 7% of tweets don't have a determined language (those marked as "und") and more than 5% are assigned a language, such as "ta", "in", "tl" or "cy", which can't be found in twitter developer help:
https://developer.twitter.com/en/docs/developer-utilities/supported-languages/api-reference/get-help-languages

Figure 1 shows the distribution of languages in the tweets captured:

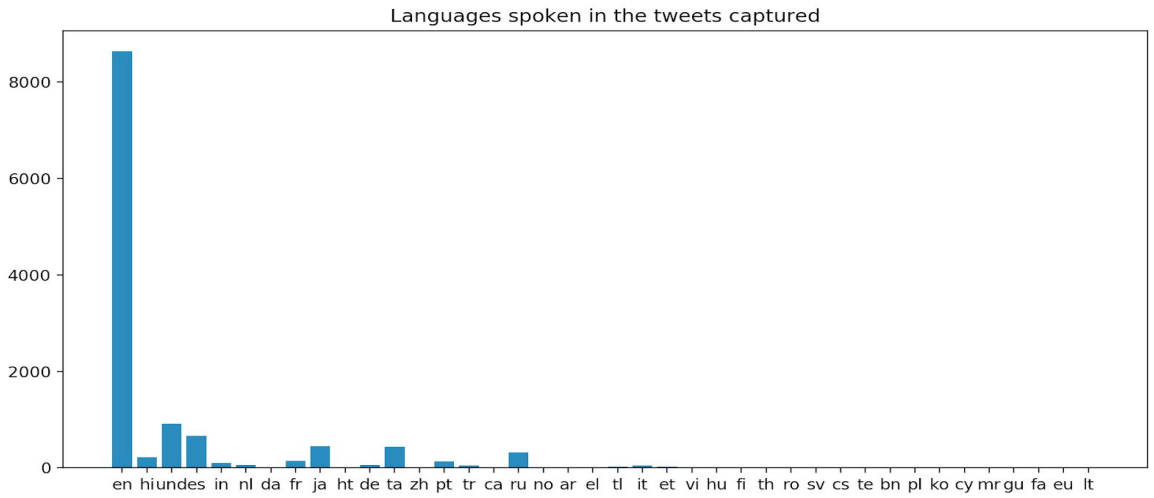

Figure 1: Distribution of languages

The numbers for each of the languages are as follows:

```
Languages: {'en': 8637, 'und': 916, 'es': 661, 'ja': 449, 'ta': 437, 'ru': 318, 'hi': 222, 'fr': 139,
'pt': 129, 'in': 99, 'nl': 60, 'de': 55, 'tr': 47, 'it': 43, 'tl': 22, 'et': 21, 'ca': 17, 'ro': 12,
'hu': 8, 'th': 7, 'ar': 6, 'el': 6, 'zh': 5, 'da': 4, 'ht': 4, 'vi': 4, 'sv': 4, 'cs': 4, 'cy': 4,
'no': 3, 'bn': 3, 'pl': 3, 'ko': 3, 'eu': 3, 'lt': 2, 'fi': 1, 'te': 1, 'mr': 1, 'gu': 1, 'fa': 1}
```

Percentage of tweets depending on how the content is generated

Using the given code, it's shown that 52.32% of the captured tweets are **original** content and 41.85% are **retweets**, which gives us still hope on people capable of writing their own stories instead of only retweeting stories that other people posted.

Assuming that **quoting** a tweet means that one has to read a person's tweet, it can be deduced that not many people are interested in what others say and they only focus on writing new content regarding what is already published in the network.

A similar assumption can be made for **replies**, as they are the lower percentage with only 2.06% of the total tweets.

Watching these percentages one might think that the behaviour of a Twitter user is that of a person who is only interested in publishing whatever information or retweet others', with no interest in discussing the different topics that other users have posted. An increase in replies would help to create a community around a topic who has common interests beyond personal ones, who is eager to learn, grow, and improve human relations.

Being the subject of tweets Yoga, I personally find interesting and useful that most of the content is original, as this is an interesting field where the point of view of different people from one subject helps others understand all better. On the other hand, I am pretty sure that most of this content is personal publicity.

Figure 2 shows the distribution of content in the tweets captured:

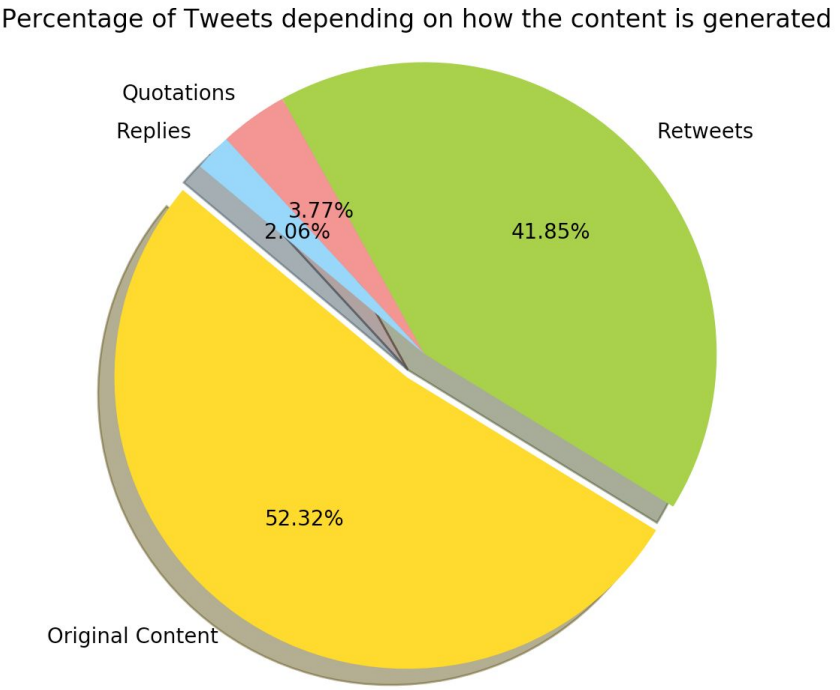

Figure 2: Distribution of content

The numbers for each type of content are as follows:

```
- Content generated via (number of tweets, from a total number of 12362:
  Original: 6468 (52.32%) / Replies: 255 (2.06%) / Quotations: 466 (3.77%) / Retweets: 5173 (41.85%)
```

Top 30 secondary hashtags

In Figure 3 the top 30 secondary hashtags are shown, excluding “yoga” as this is the main hashtag:

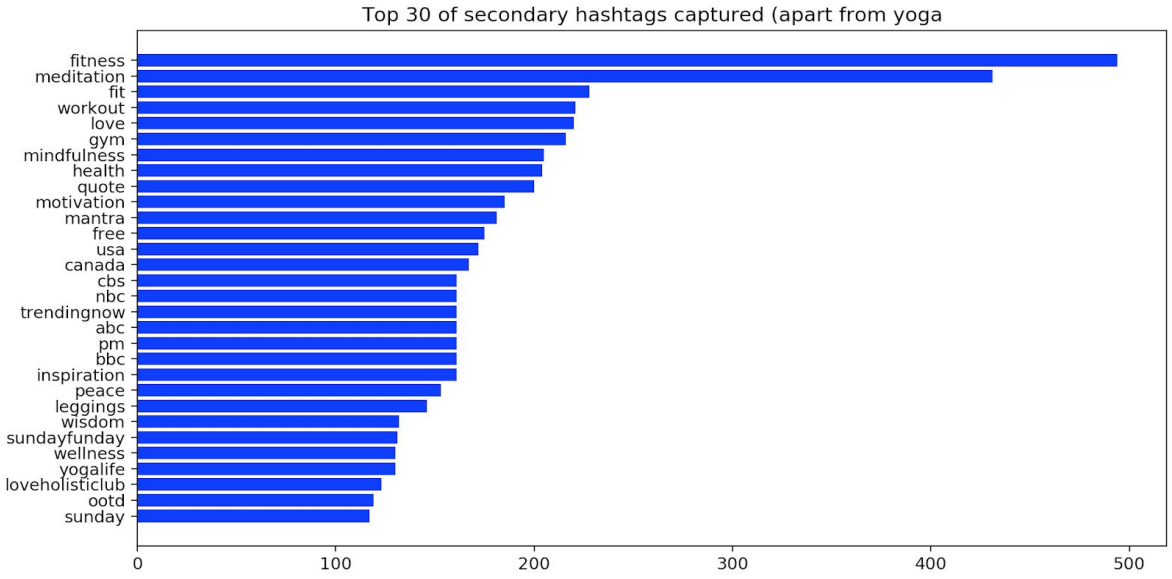

Figure 3: Top 30 secondary hashtags

What the plot shows is that yoga is mainly related to fitness, meditation and positive state of mind. Fitness and meditation have the double of occurrences than the next hashtags.

Looking at the results in more detail, the sum of hashtags related to the different topics is the following:

- fitness sum up: 1159 tweets (9.37% of the total)
- meditation sum up: 817 tweets (6.60% of the total)
- positive mind sum up: 1270 tweets (10.27% of the total)

The numbers for secondary hashtags are as follows:

```
- Top 30 of secondary hashtags:   [('fitness', 494), ('meditation', 431), ('fit', 228), ('workout',
221), ('love', 220), ('gym', 216), ('mindfulness', 205), ('health', 204), ('quote', 200),
('motivation', 185), ('mantra', 181), ('free', 175), ('usa', 172), ('canada', 167), ('cbs', 161),
('nbc', 161), ('trendingnow', 161), ('abc', 161), ('pm', 161), ('bbc', 161), ('inspiration', 161),
('peace', 153), ('leggings', 146), ('wisdom', 132), ('sundayfunday', 131), ('wellness', 130),
('yogalife', 130), ('loveholisticlub', 123), ('ootd', 119), ('sunday', 117)]
```

They show that a state of positive mind seems to be the objective of this practice, through physical exercise and meditation. Not far from reality, although between body and mind people seem to vote for the first as more important than the second.

50 most common words in twitter texts

The content of tweets can provide information about the interest of the users related to the hashtag “yoga”.

Given the text of the 12362 tweets captured, the words of each sentence have been stored and processed so that “@” and “#” characters at the beginning of words or “.” at the end have been removed. The word “yoga” has also been discarded, as is the main hashtag, and only the words that are nouns (Part Of Speech starting by “N” in NLTK python module) and doesn’t start by “http” have been finally stored in the database.

Similar to what has been seen in the previous subsection (Top 30 secondary hashtags), the topics that characterize the texts of tweets are the same as the secondary hashtags have shown. Therefore, the hashtags are like a quick summary about the text of the tweets.

The numbers for the 50 most common nouns in texts are as follows:

```
- 50 most common words in twitter texts: [('fitness', 422), ('meditation', 390), ('leggings', 315),
('love', 303), ('sunday', 296), ('workout', 295), ('health', 288), ('design', 256), ('yoga…', 219),
('day', 215), ('time', 208), ('body', 207), ('shop', 199), ('music', 191), ('way', 183), ('gym', 179),
('wellness', 173), ('mantra', 172), ('life', 169), ('today', 166), ("it's", 165), ('peace', 163),
('practice', 157), ('nikitahsolanki', 155), ('explore', 155), ('start', 155), ('pose', 154), ('mind',
154), ('shraddhadas43', 153), ('world', 147), ('morning', 147), ('motivation', 146), ('video', 143),
('inspiration', 143), ('gratitude', 143), ('class', 143), ('enuecoshop', 140), ('inspirations', 135),
('mindfulness', 133), ('moe', 133), ('something', 131), ('yoga,', 129), ('yogi', 129), ('friday',
128), ('sundayfunday', 127), ('join', 126), ('loveholisticlub', 123), ('soul', 121), ('december',
116), ('divine', 114)]
```

Figure 4 shows the 50 most common nouns in twitter texts.

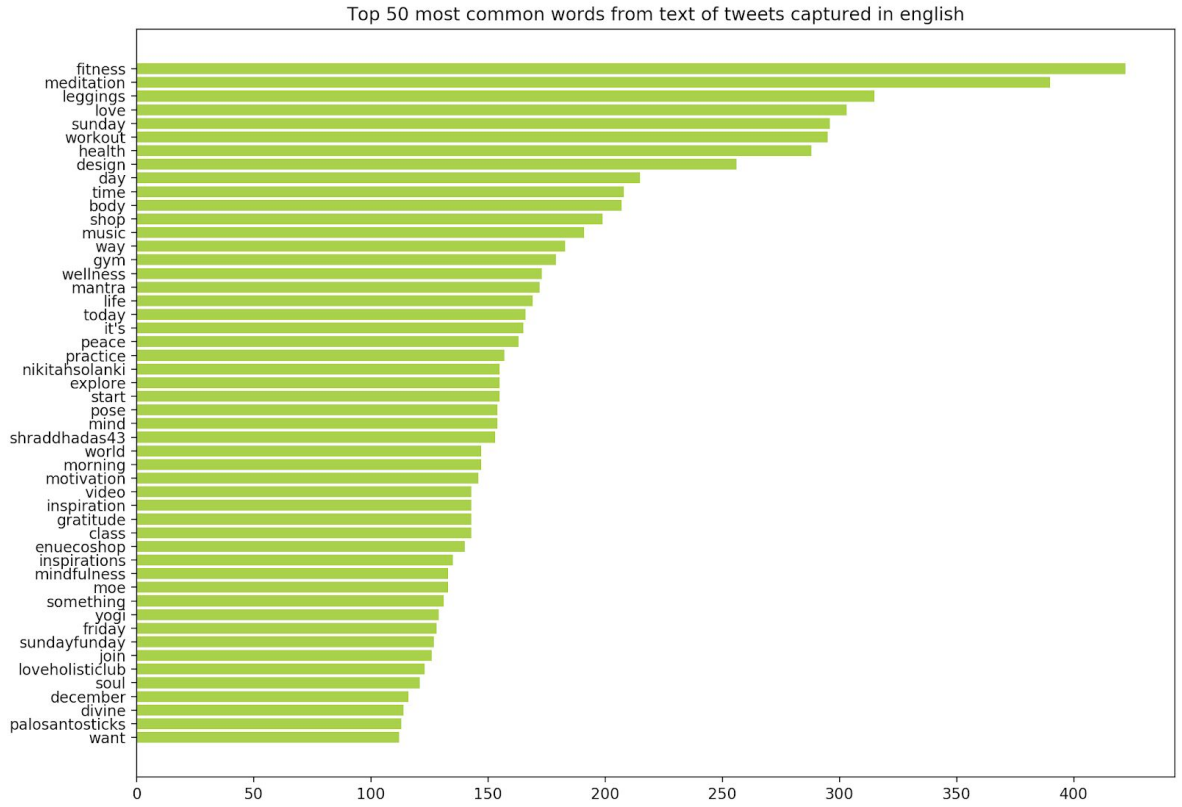

Figure 4: Top 50 most common words in texts

PART 2: Yoga articles from 'YOGAMAG' and 'HEALTH AND YOGA' websites

For the second part of this project, the following websites have been analysed:

- yogamag.net: magazine of the Bihar school of yoga (India). Its archive contains all the articles from the magazines published since 1975 to 2016 classified by year.
- healthandyoga.com: online shop with a subsite with a collection of yoga articles classified by topic.

From each of the websites all the titles of the articles have been extracted, being the total number of articles 3771; 3523 from yogamag (93.42% of the total) and 248 from healthandyoga (6.58% of the total).

The sizes of the groups of articles are very unbalanced, therefore the discussion on the topics of each one will be addressed separately instead of altogether.

yogamag.net

In Figure 5, there can be seen the number of articles published per year (between 1975 and 2016):

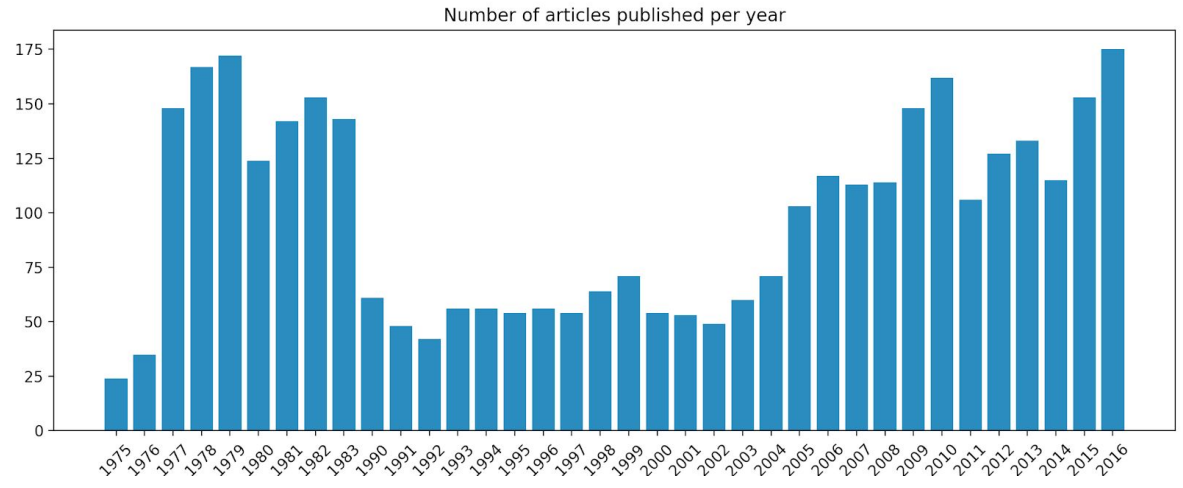

Figure 5: Number of published articles per year (yogamag)

From the counting on the names of the articles, from 1975 until 2016, it can be seen that some articles have been reedited through the years. This can be assumed as it would make no sense to publish one article more than once in the same monthly edition.

The most successful ones are shown in Figure 6 below:

Name of Article	Count
Sayings of a Paramahamsa	50
Yoga Research & Therapy	37
Yoga Research & Therapy Research reports correlated by Dr Swami...	16
Satsang at Ganga Darshan	11
Satsang with Sri Swamiji	9

Figure 6: Number of published articles per year (yogamag)

As seen in Figure 6, the articles that have been published more times are related to the sayings of the paramahamsa, who is a Hindu spiritual teacher who has become enlightened. Also Q&A (Satsang) with other teachers seem to be important issues to discuss.

A more scientific view of yoga is also treated in a lot of articles, where they make some research on different issues and how Yoga can help people to feel healthy through its practice.

As mentioned before, this journal is from India, and it can be seen that the importance of the articles coincides with the Hindu view of yoga as a philosophy, contrary the occidental view that sees it as a way to be physically healthy and, in the process, also mentally.

They live under the advice of the science of yoga, and they study not only how to understand our own mind but the effects of the state of the mind in the body.

healthandyoga.com

In figure 7, there can be seen the number of articles published per topic, where “therapy” is the topic with most of the articles. It represents the 40% of the total number of articles.

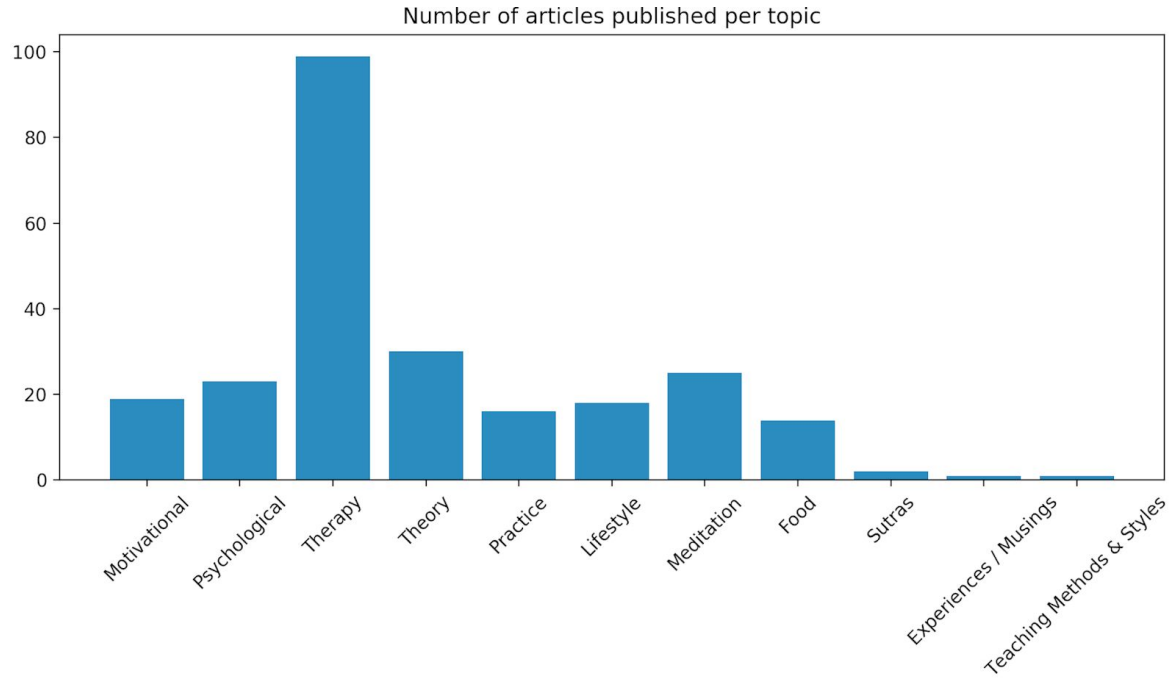

Figure 7: Number of published articles per topic (healthandyoga)

Given the shortage of articles in this website, there are seldom articles published more than once, which are not on the topic of Therapy but in Meditation, Psychological, Theory and Motivational. This is shown in Figure 8 below:

Title of the article	**Topic category**
Why Do We Sing	Theory
Meditation (Dhyana) By Shyam Mehta	Meditation
The Power Of Mantras And Malas	Psychological
Managing Negative Mental Health Through Yoga	Meditation
The 3 Step Approach To A Positive Attitude	Psychological
Understanding The Meditation Process	Meditation
Stand And Bloom	Motivational

Figure 8: Articles published more than once (healthandyoga)

As Therapy is the topic with more articles, that shows that this is the topic that people are more interested in. Some articles into this category would be “Yoga For Arthritis”, “Yoga For Digestive Disorder” or “Yoga Practice In Pregnancy” to name a few.

This is an occidental website and, as can be seen by the name of the articles, they refer to tips of yoga that people can use in they normal life to feel better, to improve the quality of their day to day. More as an aspirin to take when life overwhelms us than a proposed way of living and a way to see the world, how we act, etc, as the Hindu view proposes.

Analysis per words

To have more insight about the topics discussed in yoga articles, a count of the words that are nouns (Part Of Speech starting by “N” in NLTK python module) that don’t start with “http” or yoga have been finally stored.

In figure 9 the 20 most common words of each of the websites’ articles are shown:

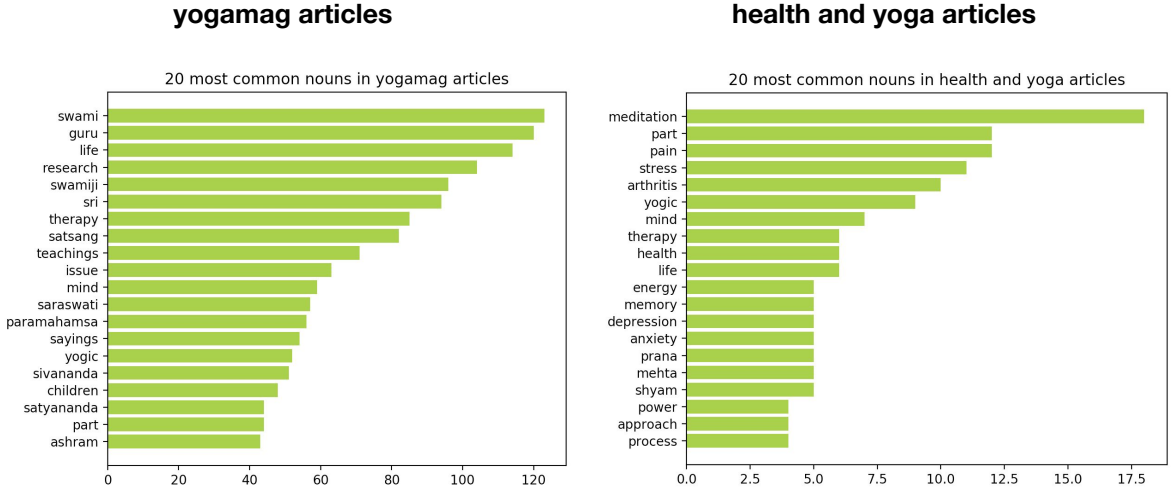

Figure 9: 20 most common words. Left: yogamag. Right: heath and yoga

Coinciding with what has been said previously, in yogamag articles (Hindu view of yoga) the words refer to philosophy and a way of living as well as the research that is done around this science.

On the other hand the words of health and yoga articles (occidental view of yoga) refer to tips to be healthy and to use against a variety of diseases that humans can suffer, though is “meditation” the word with most occurrences on these articles.

Conclusion

It has been seen that the occidental view of yoga is mostly what twitter reflects: fitness + meditation, having each of those hashtags the double of the occurrences that the next ones.

As a social media network, it is probably used more for publicity than to add content of interest, as the name of some media networks also appear in some secondary hashtags.

The fact that the original content is higher than the retweets which could show more interest in the subject than what happens with some other topics, where there are a huge amount of retweets to take part in the game of social media with any profound interest in the topic itself.

It is important to emphasise that “meditation” is the most used noun in the occidental articles, which would indicate that something written in an article is more through-out that some post in a twitter account about the same topic. This is a step closer to the view of the Hindu magazine that shows yoga how it really is.

Needed to say that the people behind the yoga articles, being those occidental or Hindu, have more skills in the topic than most of the people that post in twitter, which explains why the content is different from articles to twitter posts.

Git Repository

In https://github.com/ovalls/mai_bda you can find the code for this project (folder “Project”) along with some other scripts, files and plots:

- `articles.py:` script to get and analyse the titles of the articles from two yoga websites.
- `analize_yoga.py`: script to analyse twitters from the hashtag #yoga
- `stream_no_keys`: script to get #yoga tweets and store them in a Mongo Database.
- `BDA_Project_OlgaValls.pdf`: This report.