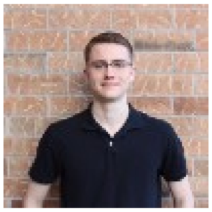
Exploring 16000 Frontiers in Neuroscience Articles

towardsdatascience.com/exploring-16000-frontiers-in-neuroscience-articles-7f362de76ba3

June 15, 2019



Recently I've been working a lot on voice apps. One thing that really sticks out for me is how natural a conversational interface is. If the technology advances enough it will allow us to complete tasks more effortlessly since we can simply say what it is we want to do. However, no matter how efficiently a device is surfacing information, your ability to digest information will always be limited by the rate at which you can read, listen, or watch.

This is what makes the idea of advanced Brain Computer Interfaces (BCIs) so great. They would allow people to learn, retain knowledge, and communicate more efficiently than ever before. Elon Musk has been talking about how his mysterious Neuralink startup is developing high bandwidth BCIs to connect humans and computers. The idea behind Neuralink is that if the brain can be both interpreted and manipulated in real time you could “extend” the brain. This could add a new layer of capability, understanding, and communication. People will be able to transfer ideas directly to another person without needing to communicate with written / verbal language. However, BCIs have been researched for a while now, with many applications focusing on helping people suffering from paralysis.

The recent hype behind Neuralink is the inspiration behind this analysis. I wanted to explore what are the some of the trends in brain computer interface research. So after setting my goals I got ready to spend the day battling Pandas and Matplotlib to see if I can glean out any interesting information from my dataset.

Preparing the Dataset

Frontiers Media is an academic publisher of peer-reviewed open access scientific journals. They have numerous journals in science, technology, and medicine. For my analysis I used articles from the Frontiers in Neuroscience journal series which contains some of the world’s top neuroscience journals in terms of influence and quality. Some of these journals include Frontiers in Neuroscience, Frontiers in Neuroinformatics, and Frontiers in Cellular Neuroscience.

My first step was collecting the data. In this case I wrote a Python script to scrape all of the neuroscience articles. The web scraper saved the article’s text as well as their metadata. To collect the URLs of all the articles, I used a bit of regex to parse the sitemap.xml file for the URLs in the neuroscience series. Once my web scraper finish running, my dataset contained 15803 articles stored in a JSON format. These articles came with a rich array of metadata including publication countries, authors, citation dates, and keywords.

Converting the JSON dataset to a CSV is no problem with Pandas. You can simply use the normal dataframe constructor if the JSON object is flattened out (only one level of keys). However looking a bit closer reveals that there are empty dict objects in fields that were missing a value. Here’s a neat trick you can use to replace these empty dicts with NaN:

```
df = df.mask(df.applymap(str).eq(‘{}’))
```

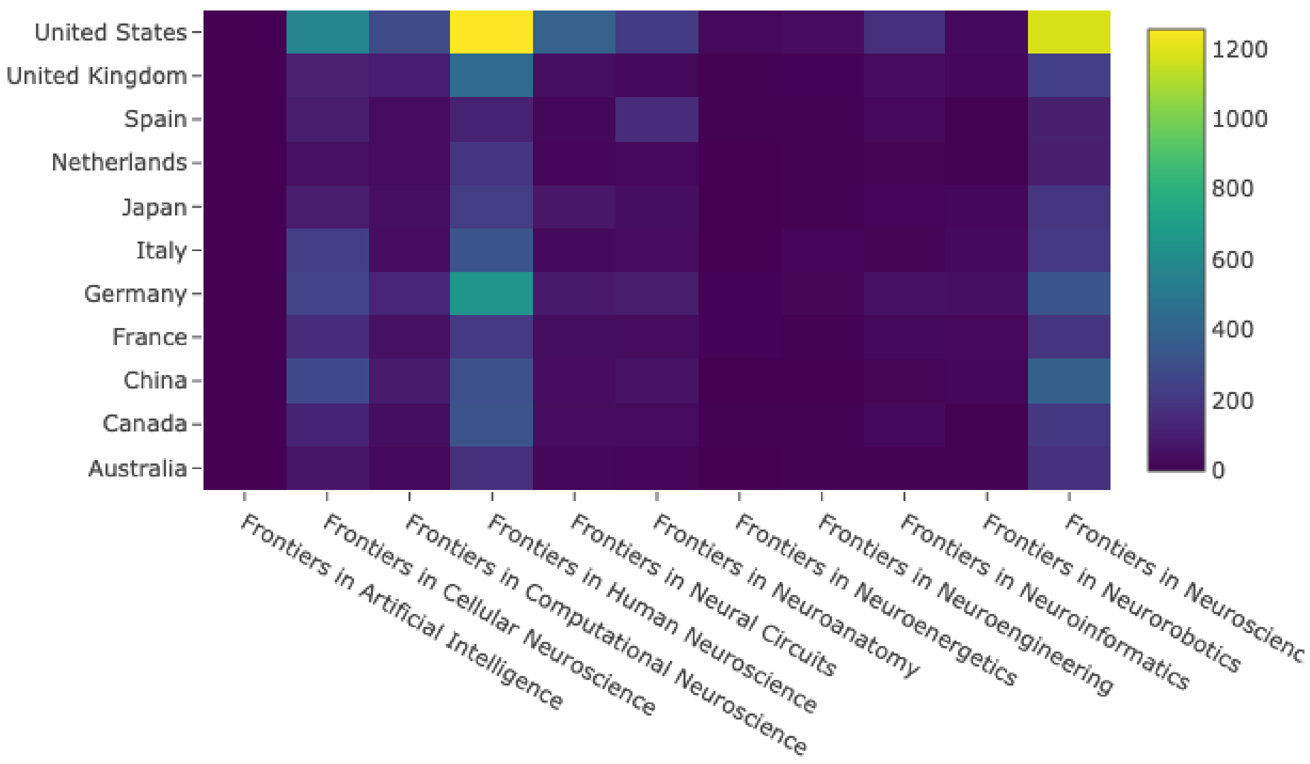
Now that everything looks good, let’s call df.to_csv(), upload the dataset to Kaggle, and do some exploratory data analysis.

Check out the Kernel: https://www.kaggle.com/markoarezina/frontiers-in-neuroscience-article-eda

You can find the dataset on Kaggle: https://www.kaggle.com/markoarezina/frontiers-in-neuroscience-articles

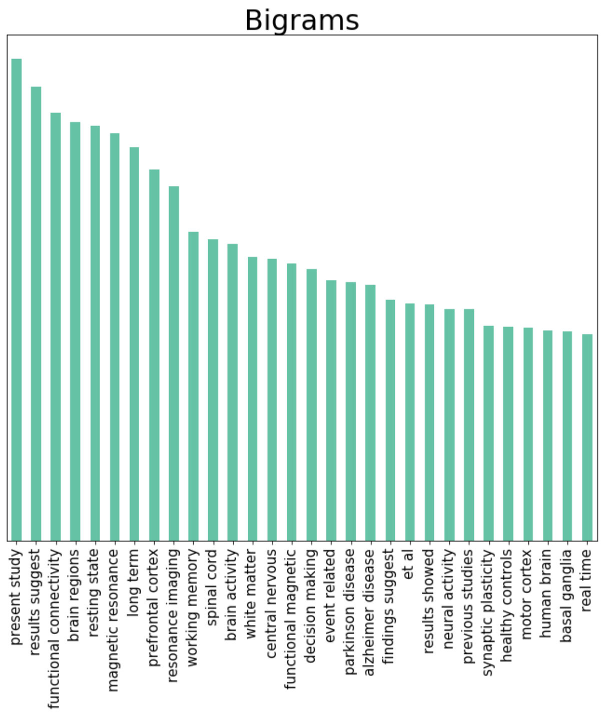
Getting to Know the Data

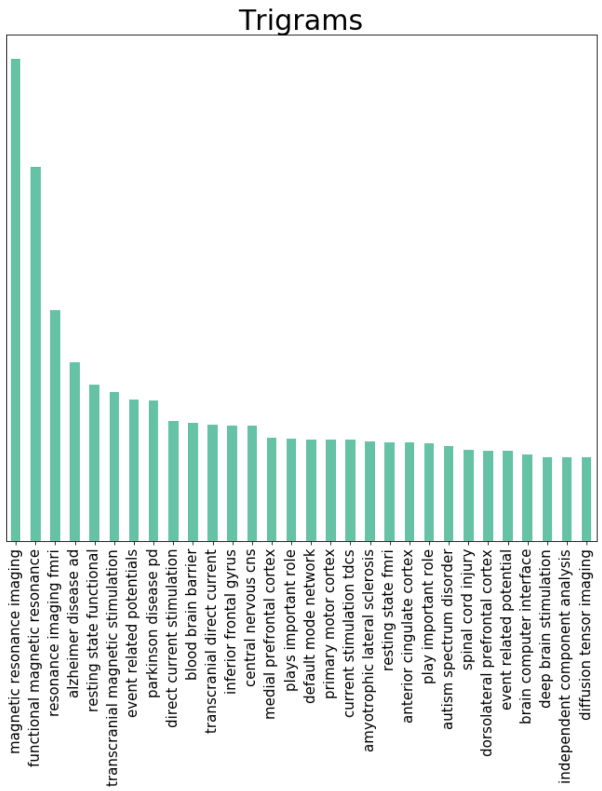
I wanted to get a feel of the types of articles in the dataset before taking a deep dive into BCI specific research. Apart from the journal text, the dataset also contained a wide range of interesting metadata. This included useful fields like the “citation_journal_title“ which specified the article’s specific neuroscience journal. For example, Frontiers in Cellular Neuroscience. Since the metadata also contained the citation country, I thought a heat map of countries and neuroscience journals for number of articles might be a good way to get an overview of the the dataset.



From the heat map you can see some of the largest journals in the series are Frontiers in Cellular Neuroscience, and Frontiers in Neuroscience, and Frontiers in Human Neuroscience.

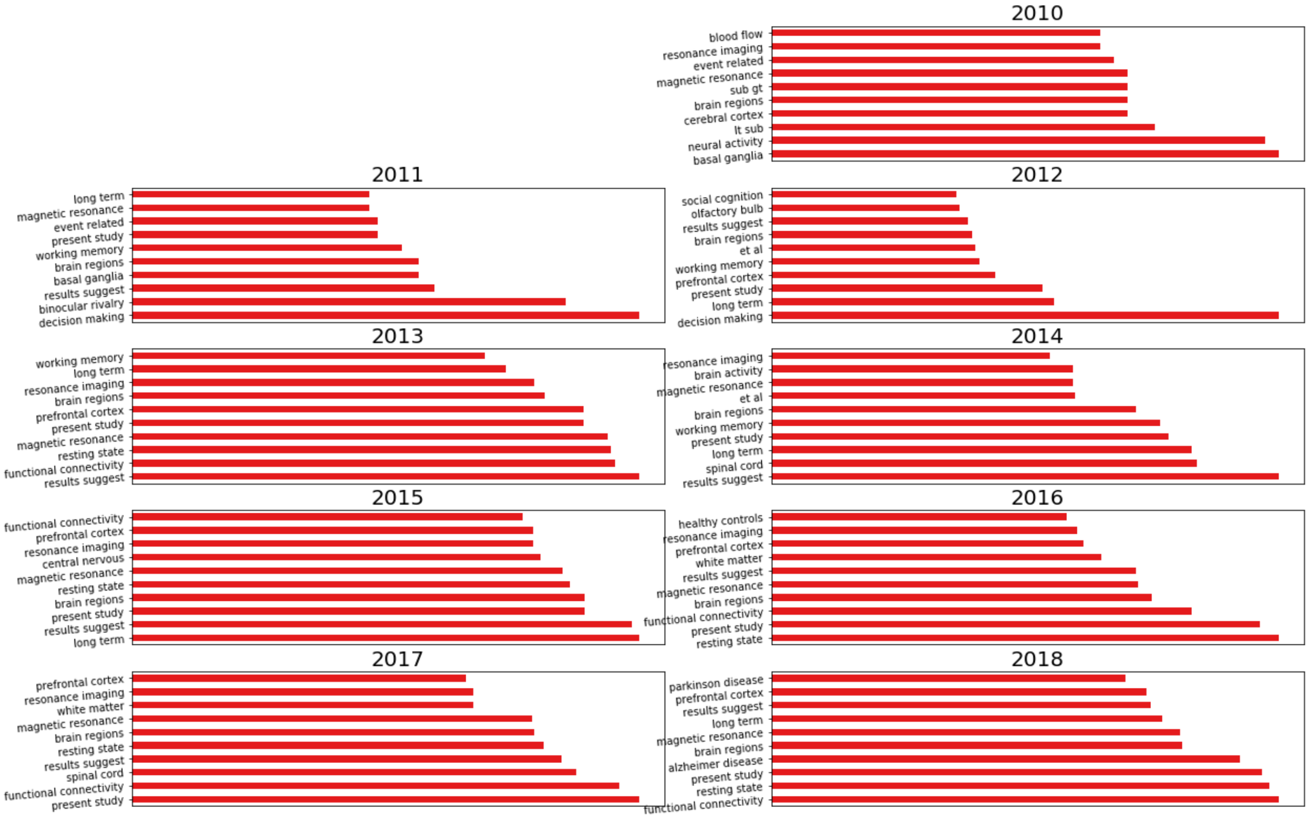
Next, I wanted to explore some of the common themes in the articles. To do this I found the most common unigrams, bigrams, and trigrams of the citation abstracts. The following plot shows some common themes.



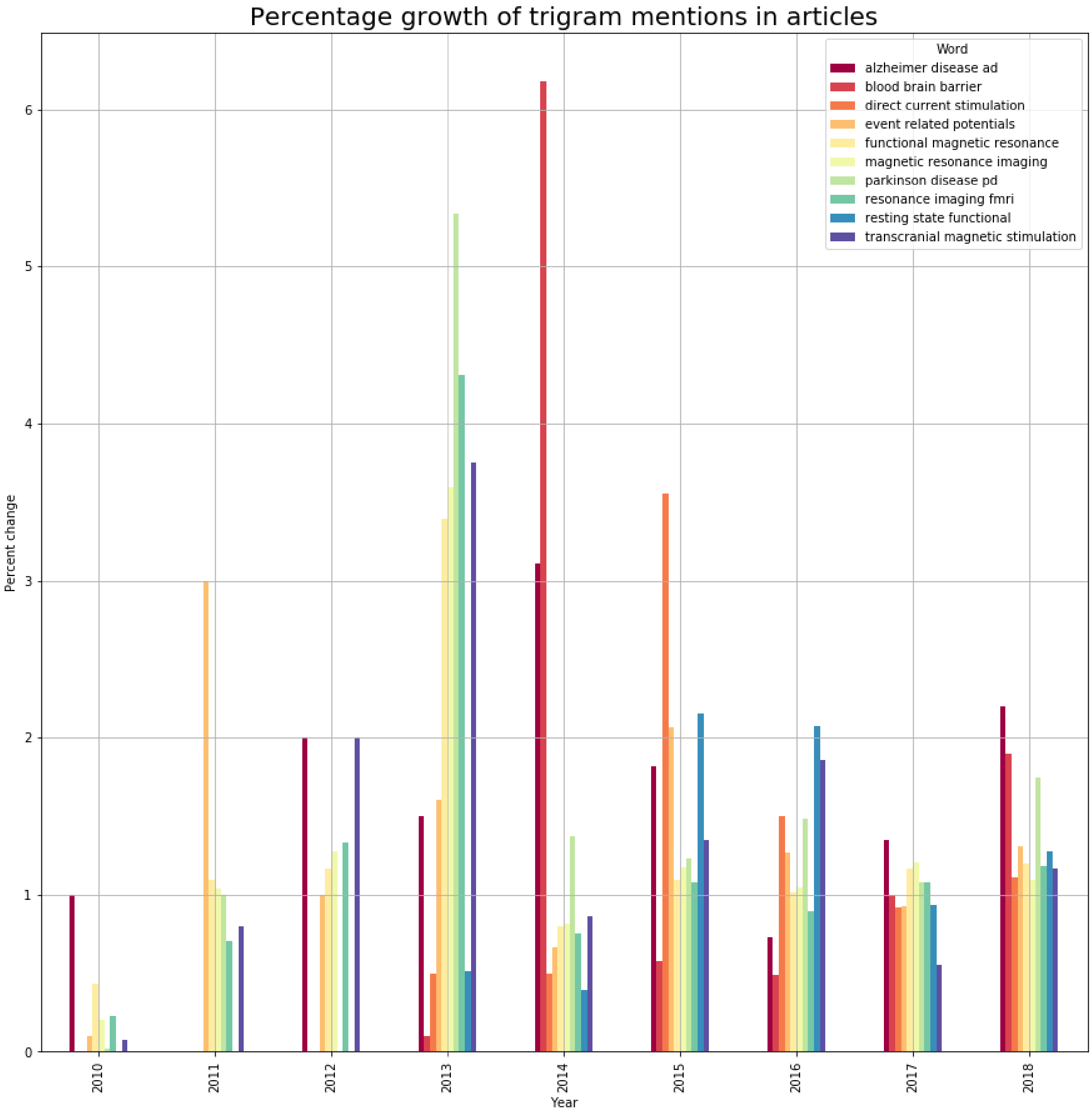


The unigrams did not reveal much, with unsurprisingly the most common word being “brain”. The most common trigrams are related to imaging which is essential for studying the brain . Magnetic resonance imaging uses strong magnetic fields and radio waves to generate images of the organs in the body. Functional magnetic resonance refers to a method for measuring brain activity by detecting changes associated with blood flow.

I also though it might be a good idea to analyze the popularity of bigrams over time. This reveals some key concepts like magnetic resonance which is a top bigram in all the years. It also suggests how the research focus shifted over the years.



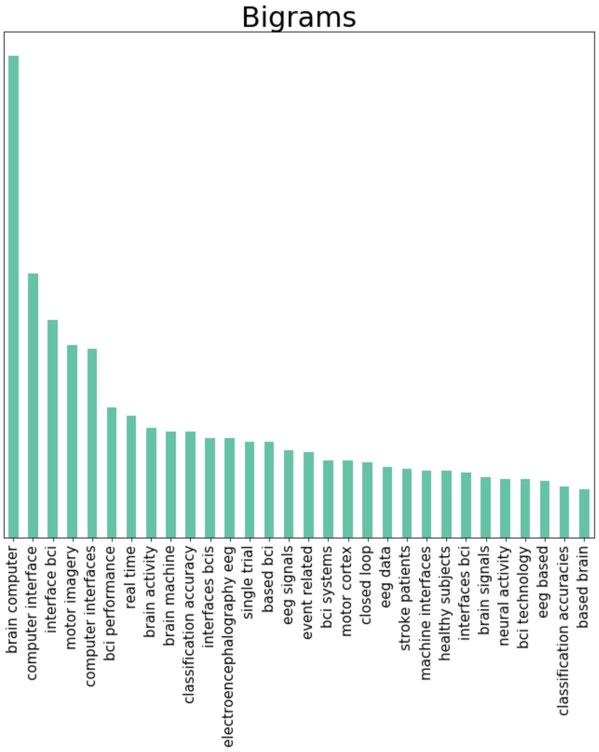
Plotting the percentage growth of trigram mentions over time, shows spikes in certain keywords as there is an influx of research papers in a certain year.

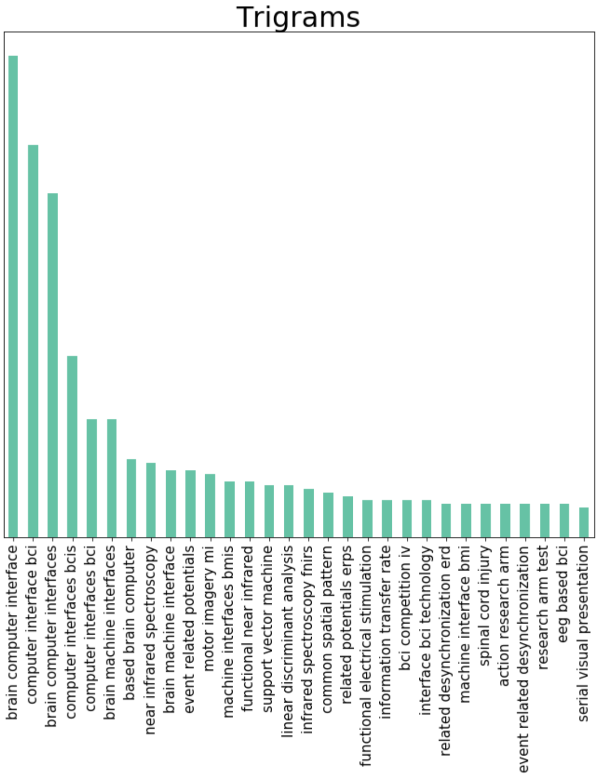


BCI Specific Research

To get started with analyzing some of the themes specific to BCIs, I queried articles that mentioned the phrases brain computer / machine interface in the abstract. This gives us 332 articles.

Running a similar analysis of bigrams and trigrams reveals some common themes in the BCI specific articles.



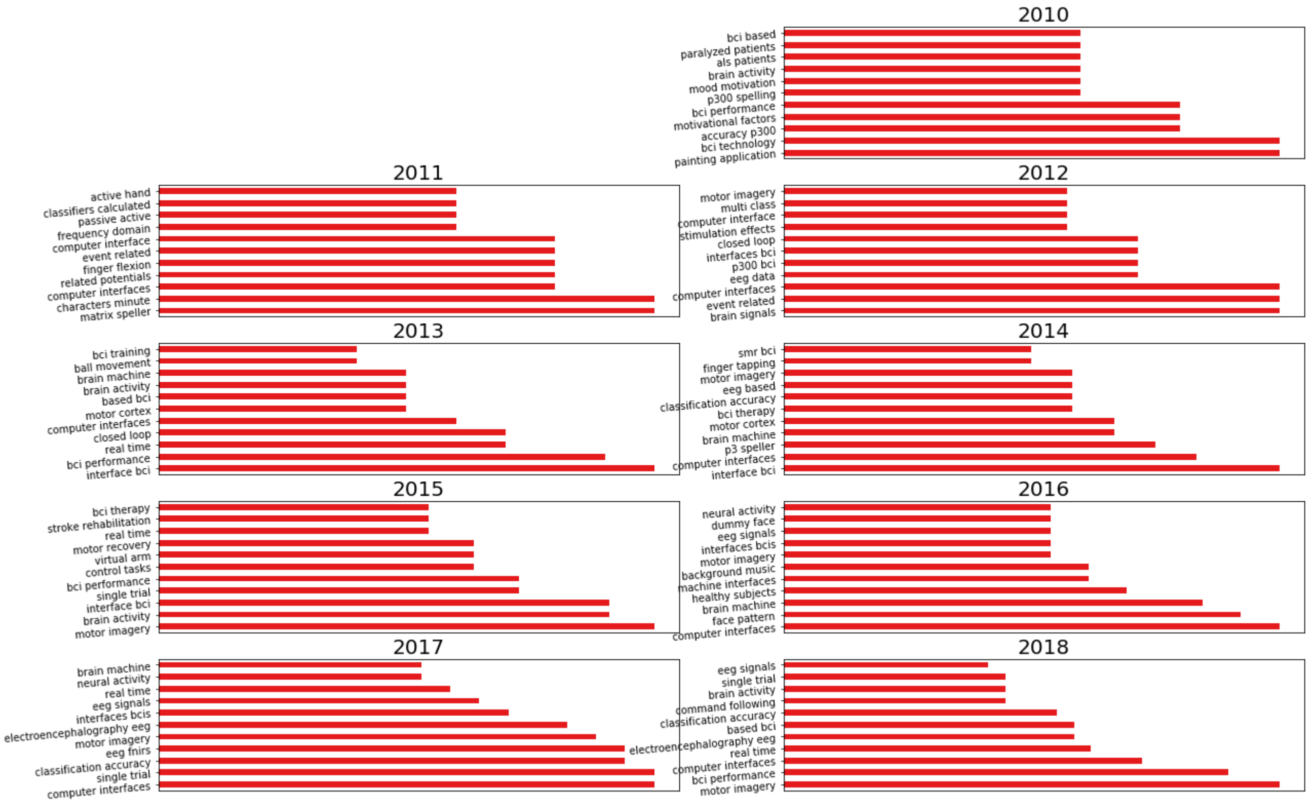


Here are some of the most common BCI specific terms and their explanations:

motor imagery: is one of the standard concepts of BCI, in that the user can generate induced activity by imagining motor movements. Users may need several training sessions before they learn how to generate desired brain activity and deliver the intended command. A typical training protocol for such BCIs includes execution of a motor imagery task by the user, followed by moving object on a computer screen.

“Support vector machine” / “Classification accuracy”: These phrases are appearing in the context of classifying the Motor Imagery EEG Signals. Support Vector Machines can be used to classify Motor Imagery EEG Signals to determine which action the user is trying to take.

Plotting the bigrams over time reveals the shift in focus over the years.



Some interesting patterns include the focus on the p300 wave along with its application in a BCI for spelling. There is also an increase in focus on electroencephalography and classifying these signals in the years 2017 and 2018.

Some interesting things to try in the future include comparing the research themes between the different journals in the series. It would also be interesting to explore more of the data which includes institutions, keywords, and references. I don’t know much about neuroscience and I found this analysis to be a great way for getting an overview of some of the research areas in neuroscience with a focus on BCIs. I’d love to hear what you would do better or if there is something I could correct.

I develop Google Assistant and Amazon Alexa skills at Vecgraph. Currently, we are looking to partner with a few brands and develop their voice integration. Since we are looking to collaborate on a case study, we are offering a discounted rate for the design and development of your voice app. This opportunity is first come first serve. Let me know what you’re looking to create! marko.arezina@vecgraph.com