Name : Ioan-Radu Bocu
Student number : s1104958

Project Report

Description and Motivation

The main goal of this project is to compare the corresponding number of occurrences for 2 types of 2 diets : ketogenic diet and vegan diet. This project aims to provide insights about how frequent these terms are mentioned on the Internet and also to determine which diet was more popular till the last update of the cluster (2021). This comparison can be further used in relation with current statistics or older statistics based on this trend in order to observe the dietary preferences of the people over the time.

Both keto and vegan diet have drawn significant attention in recent years. The keto diet implies eating low amount of carbs and a huge amount of fat (preferring fat in the meat over the carbs). The diet has a significant potential for weight loss. On the other hand, the vegan diet , which excludes all animal products, is often chosen for health, ethical and environmental considerations. Analyzing the frequency of these types of diet could indicate which method is more popular and effective.

Process

To begin with, I defined a number of functions that would help in achieving the goal of the project. The primary concern was processing WARC files, parsing the HTML content , and searching for references to the keywords “keto” and “vegan” within the text body of each web page. The are the steps that were done:

1. **Parsing HTML content** : The Jsoup library was used to parse the HTML content of each web page contained in the WARC files.

2. **Filtering Relevant Records :** We had to assure that only records with non-empty URLs and bodies are processed.

3. **Checking the language** : I did this, as I preferred getting results only from sites in English.

4. **Counting the corresponding references** : In this step, I counted the number of occurrences f each corresponding word using : val ketoPattern = "(?i)\\bketo\\b".r val veganPattern = “(?i)\ \bvegan\\b".r for pattern matching. My search was done in a case-insensitive approach in order to obtain relevant results for the analysis,

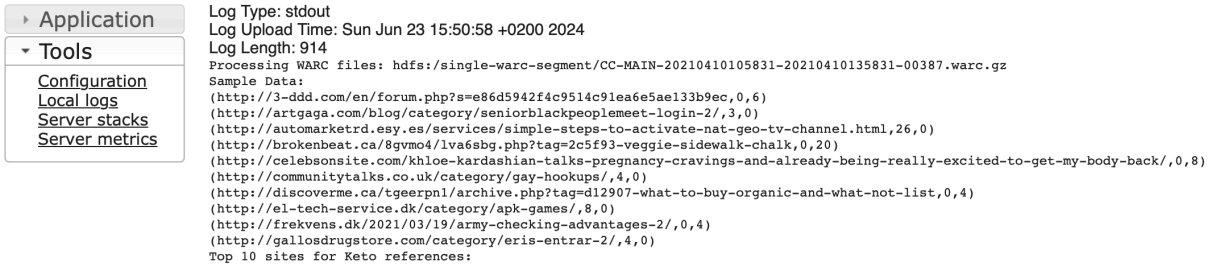
5. **Calculating Standardized counts :** I have divided these counts by the total word count of each page.

6. **Creating DataFrame** : Our results were stored in a DataFrame for further analysis.

Testing and Challenges

One of the first problem was correctly identifying and extracting the body content of the websites. Overcoming this problem is essential in order to obtain accurate keyword counting. The problem

was : variations in HTML structures (the HTML structures can vary significantly across different websites) and presence of non-informative content. Some websites may have the main content wrapped within multiple nested or <section> tags, whereas other sites could used completely different HTML elements. Besides this, we also need to be careful to non-relevant content such as ads, navigation bars, and footers which could mislead the counting if we do not do a proper filtering.



Another problem that I encountered is given by the screenshot from above. In the case above, I was doing some debugging, cause my submitted tasks were constantly failing. The debugging was done in the following way : I did some intermediary printings ; printing the sample data and the WARC files being processed in order to ensure myself that the code was not messing up the elementary steps.

I solved the issue from above in the following way : the error handling was improved by wrapping the parsing and keyword counting logic inside a try-catch block. This modification is quite important as it ensures that any exception during the processing of the WARC files are caught , preventing the forced execution stop of the task.

I also increased the memory allocation for both the driver and executors in order to be sure that large datasets can be handled properly. Off-heap memory was enabled in order to provide additional memory for processing.

Results and Analysis

I have run my code on cluster on 2 segments and 10 segments, in order to observe how the results are changed when increasing the number of files being processed.

2 WarcFiles execution

```
Log Length: 2581
Processing WARC files: hdfs:/single-warc-segment/CC-MAIN-20210410105831-20210410135831-00000.warc.gz,hdfs:/single-warc-segment/CC-MAIN-20210410105831-20210410135831-00001.warc.gz
Filtering WARC records completed.
Processed records count: 581

########## Start ##########
Top 10 webpages in WARC files mentioning 'keto':
URL: https://community.buzrush.com/tag/packaging-box-manufacturers/, Keto Count: 1387
URL: https://community.buzrush.com/profile/john/points/, Keto Count: 1387
URL: https://www.oprebatph.org/ketodietreview/cheap-plan-custom-keto-diet-buyback/, Keto Count: 60
URL: https://s3-us-west-1.amazonaws.com/food-a35785caaa9c49abafc299bf3b68534c/page-63680685030012062 7.html, Keto Count: 34
URL: https://jenniferbanz.com/category/recipes/dinners/page/3, Keto Count: 28
URL: https://jenniferbanz.com/category/recipes/low-carb-30-minute-meals/page/2, Keto Count: 24
URL: https://fastweightlossdiary.com/lose-weight-fast/9-cute-pieces-to-inspire-everyone-you-see-to-vote/, Keto Count: 24
URL: http://www.kabarkhusus.com/2019/09/7-makanan-untuk-menu-diet-keto.html, Keto Count: 16
URL: http://www.ymdb.ca/compounds/YMDB00153?kegg_reactions=2, Keto Count: 15
URL: http://www.city-star.org/guestbook/en/6182, Keto Count: 13

Top 10 webpages in WARC files mentioning 'vegan':
URL: https://fiordizucca.blogspot.com/2006/01/pasticcini-glassati-di-anice-e-pinoli.html?showComment=1137360360000, Vegan Count: 37
URL: https://www.trendeing.com/vegan-eats/vegan-thanksgiving-made-easy/, Vegan Count: 32
URL: https://yourbagheaven.com/collections/holiday-shop, Vegan Count: 24
URL: http://www.adasvegan.com/project/vegan-philly-sandwich/, Vegan Count: 23
URL: https://cbjspotlight.co.uk/2017/02/02/nottingham-embraces-veganuary-campaign/, Vegan Count: 21
URL: https://peanutbutterandjilly.com/recipes/20-minute-peanut-chocolate-chip-cookies/, Vegan Count: 20
URL: https://nom-noms.de/en/tag/bagels-en/, Vegan Count: 18
URL: https://fittingintovegan.com/2016/12/13/vegan-and-gluten-free-chai-spiced-sweet-potato-casserole/, Vegan Count: 14
URL: https://www.yumbles.com/gourmet-spice-company/gourmet-salt-and-pepper.html, Vegan Count: 13
URL: http://sshskola.sk/ryrqwtl/41565f-la-fiorentina%2C-bristol-menu, Vegan Count: 12

Results Summary:
Term       Total References     Percentage
Keto       3280                 71.18%
Vegan      1328                 28.82%
########## Ending ############
```

10 WarcFiles execution

```
Log Type: stdout
Log Upload Time: Mon Jun 24 02:34:16 +0200 2024
Log Length: 2794
Processing WARC files: hdfs:/single-warc-segment/CC-MAIN-20210410105831-20210410135831-00000.warc.gz,hdfs:/single-warc-segment/CC-MAIN-20210410105831-20210410135831-00001.warc.gz,h
Filtering WARC records completed.
Processed records count: 2865

########## Start ##########
Top 10 webpages in WARC files mentioning 'keto':
URL: https://community.buzrush.com/tag/active-fitness-keto-blend-reviews/, Keto Count: 1395
URL: https://community.buzrush.com/tag/keto-slim-t3-reviews/, Keto Count: 1389
URL: https://community.buzrush.com/tag/packaging-box-manufacturers/, Keto Count: 1387
URL: https://community.buzrush.com/profile/john/points/, Keto Count: 1387
URL: https://www.ketodietloseweightfast.com/keto-genie-diet/, Keto Count: 175
URL: https://www.ketodietloseweightfast.com/keto-diet-order/, Keto Count: 175
URL: https://www.ketodietloseweightfast.com/shakara-keto-diet/, Keto Count: 173
URL: https://www.oprebatph.org/ketodietreview/best-mid-priced-plan-custom-keto-diet/, Keto Count: 118
URL: https://dzp.uw.edu.pl/?ketodiet=carrie-underwood-keto-diet-pill_Ketogenic, Keto Count: 110
URL: https://www.tannoshealth.com/tag/labrada-net-worth/, Keto Count: 100

Top 10 webpages in WARC files mentioning 'vegan':
URL: https://stunningbathrooms.co.za/tzwo5/vegan-candy-australia-28eb02, Vegan Count: 161
URL: https://deniseminger.com/2010/09/08/brand-spankin-new-study-are-low-carb-meat-eaters-in-trouble/?replytocom=5011, Vegan Count: 75
URL: https://ecosistent.com/93hb9nd/dc3c61-vegan-date-cake, Vegan Count: 52
URL: https://deniseminger.com/about/?replytocom=164433, Vegan Count: 51
URL: http://healthmerit.net/wild-garlic/, Vegan Count: 49
URL: http://healthmerit.net/2017/06/, Vegan Count: 46
URL: https://eatingvibrantly.com/recipes/raw-vegan-almond-milk/, Vegan Count: 45
URL: https://prijatelji-zivotinja.hr/index.en.php?id=677, Vegan Count: 43
URL: https://vegetarianculinarian.com/, Vegan Count: 41
URL: https://www.fragrantvanilla.com/chickpea-cutlets-and-mushroom-gravy/, Vegan Count: 41

Results Summary:
Term       Total References     Percentage
Keto       9010                 56.59%
Vegan      6912                 43.41%
########## Ending ##########
```

1 WarcFile execution

```
Log Length: 2278
Processing WARC files: hdfs:/single-warc-segment/CC-MAIN-20210410105831-20210410135831-00000.warc.gz
Filtering WARC records completed.
Processed records count: 284

########## Start ##########
Top 10 webpages in WARC files mentioning 'keto':
URL: https://community.buzrush.com/tag/packaging-box-manufacturers/, Keto Count: 1387
URL: https://community.buzrush.com/profile/john/points/, Keto Count: 1387
URL: https://www.oprebatph.org/ketodietreview/cheap-plan-custom-keto-diet-buyback/, Keto Count: 60
URL: https://jenniferbanz.com/category/recipes/dinners/page/3, Keto Count: 28
URL: https://fastweightlossdiary.com/lose-weight-fast/9-cute-pieces-to-inspire-everyone-you-see-to-vote/, Keto Count: 24
URL: http://www.kabarkhusus.com/2019/09/7-makanan-untuk-menu-diet-keto.html, Keto Count: 16
URL: http://www.ymdb.ca/compounds/YMDB00153?kegg_reactions=2, Keto Count: 15
URL: http://www.city-star.org/guestbook/en/6182, Keto Count: 13
URL: https://rupier.se/ccyom/35bbd4-dessert-for-cookout, Keto Count: 10
URL: https://dzp.uw.edu.pl/?ketodiet=weight-loss-tracking_Keto, Keto Count: 9

Top 10 webpages in WARC files mentioning 'vegan':
URL: https://fiordizucca.blogspot.com/2006/01/pasticcini-glassati-di-anice-e-pinoli.html?showComment=1137360360000, Vegan Count: 37
URL: https://www.trendeing.com/vegan-eats/vegan-thanksgiving-made-easy/, Vegan Count: 32
URL: https://cbjspotlight.co.uk/2017/02/02/nottingham-embraces-veganuary-campaign/, Vegan Count: 21
URL: https://peanutbutterandjilly.com/recipes/20-minute-peanut-chocolate-chip-cookies/, Vegan Count: 20
URL: https://nom-noms.de/en/tag/bagels-en/, Vegan Count: 18
URL: https://fittingintovegan.com/2016/12/13/vegan-and-gluten-free-chai-spiced-sweet-potato-casserole/, Vegan Count: 14
URL: https://www.formx.eu/special-make-up/skin-illustrator-mouth-fx/index.php, Vegan Count: 12
URL: https://www.eupedia.com/forum/threads/3232-Are-you-vegetarian/page11?p=441582, Vegan Count: 11
URL: https://mega-nutrition.co.uk/product-tag/740985271162/, Vegan Count: 11
URL: https://www.veganhaven.co.uk/product/animals-homies-unisex-t-shirt/, Vegan Count: 10

Results Summary:
Term       Total References     Percentage
Keto       3043                 81.34%
Vegan      698                  18.66%
########## Ending ##########
```

The following metrics were used : total number of references, percentage of occurring in the total number of occurrences. Also the number of occurrences of both “keto” and “vegan” per website were computed.

As we can see that as the number of WarcFiles increases, we can see that the percentages of occurrences for both words balance. Using 1 WarcFile lead to a percentage of 81.34% (keto), when using 2 WarcFiles we obtained a percentage 71.18%(keto) and , finally using 10 WarcFiles , a percentage of 56.59% was obtained. We can also see this in the case of “vegan”. Also, the websites obtained for each category differ across different warcFiles . As expected, the number of occurrences increased with the number of warcFiles. It may be the case that some sites are more specialized than others, so there can be a preference for a certain word (that is found in the body).

Conclusion

The analysis of the projects , together with the increasing number of files being processed provided a clear conclusion about the current trends in people's diet. A future goal would be to provide more scalability, meaning running the code on a higher number of WarcFiles: 25,50. Then what can also be done is running the code on multiple segments and eventually the entire cluster. By doing this, we could obtain an even deeper understanding of what we already have.