**Pre-Processing Specification**

**Data Collection**

**Input:** text file consisting of all highlighted risky statements made from data collection process

**Contracts Reviewed (All privacy policies):** Asda, EE, Lycamobile, GiffGaff, O2, Tesco, Three, Virgin Media

**How contracts were reviewed:** All members of the group reviewed 2 contracts each, highlighting sentences that were deemed risky under the Terms and Conditions Guidelines created.

**Guidelines:**

1. **Collecting your device fingerprint**
   1. "We collect device-specific information (such as your hardware model, operating system version, unique device identifiers, and mobile network information including phone number). Google may associate your device identifiers or phone number with your Google Account."
2. **Deleting an account or content may result in information being retained**
   1. “Youtube may retain, but not display, distribute or perform, services copies of your content that you have removed or deleted. ”
3. **Account Deletion or Data Deletion can be authorised at any time for any reason**
   1. “GitHub, in its sole discretion, has the right to suspend or terminate your account and refuse any and all current or future use of the Service, or any other GitHub service, for any reason at any time. Such termination of the Service will result in the deactivation or deletion of your Account or your access to your Account, and the forfeiture and relinquishment of all Content in your Account. GitHub reserves the right to refuse service to anyone for any reason at any time. ”
4. **Indemnification from claims related to your account or your content**
   1. You agree to indemnify SoundCloud from all claims resulting from any third party for infringement of copyright or invasion of privacy related to your content or any activity related to your account.
5. **Ignoring the Do Not Track header**
   1. We do not support the Do Not Track browser option.
6. **Ability to change Terms and conditions at any time without notifying the user**
   1. They can change the terms of service any time they see fit, even without notification to the user. Your use of the service supposedly constitutes acceptance of the changes in the terms.
7. **Services tracking you on other websites**
   1. This service uses cookies to track you even if you are not interacting with them directly. Amazon for instance, use [cookies](http://www.amazon.com/gp/help/customer/display.html/ref=hp_468496_cookies?nodeId=468496&#cookies) to track your device and serve targeted advertisements on other websites (Amazon associates, websites using Amazon Checkout).
8. **Services requiring broad use of permissions**
   1. When installing the Facebook app on an Android phone, it allows access to the audio record path and to take pictures with the camera. This allows the application at any time to collect images the camera is seeing.
9. **Generally being careless with personal data**
   1. **Use of personal data in ads**
      1. using personal information for targeted advertising
   2. **Used to track you**
      1. E.g. geographical information about your location
   3. **Storing personal data anywhere in the world**
   4. **Distributing your personal information to third parties**
      1. Facebook give personal details to 3rd Party by default
10. **Allowing third parties to target you**
    1. Although Amazon does not share your personal information to advertisers and other third-parties that display Amazon's ads, by default they enable third-party advertises to target you (IP, cookies or “other features”)

**Implementation**

**Programming language used:** Python

**Software package used for Stemming (Snowball), Stopwords, Duplicates:** Natural Language Toolkit within Python

**How class variables were created:** When reviewing contracts, all sentences that were deemed and highlighted risky were given an identifier (0cf333). This would allow the script to determine what sentences were risky or not when converting the information into the dataset. Within the dataset risky sentences were represented with identifiers Yes or No based on being risky or not. Each row in the dataset was a representation of a sentence in the contracts.

Risky sentence:



Non-risky sentence:



**Other features involved:**

* Exclude words: removing specific words from our dataset manually
* Removing numbers: removing numbers from our dataset manually using the strip function
* Removing punctuation: removing punctuation from our dataset manually using the strip function
* Removing empty elements: removing spaces created after the stemming process has taken place. For example, after every word that has been stemmed, the program created a space. This code allowed us to remove these spaces from the dataset.

* Sorting: sorting all header words in order that we chose i.e. alphabetically A-Z

**Output:** Dataset in the format of CSV. This holds header words (all words generated from the reviewed contracts and use of stemming, removal of stop words and duplicates), class variable (Yes or No, whether the header word is risky or not, header occurrences (0 or 1 representing whether the word has appeared in the specific sentence).