**Feedback for WS3**

We received 14 submissions for the third worksheet. Here is some feedback upon the review of your submissions:

**[E1]:**

In the first example, we want you to think about how to create gold data which can be used for training the systems solving NLP tasks. When you work with real language data, it is easily observable that even the simple tasks similar to the one given [E1a] becomes complicated. One gold standard to ensure the quality of the annotations is to prepare an annotation guideline for the human annotators. The guideline should be easily understandable by the new annotators and comprehensive enough to address both standard and challenging cases in the data, as much as possible.  
The annotation guidelines are usually built incrementally in big annotation projects. The initially prepared draft is enriched with suggested strategies for the standard examples, peculiar cases, ambiguities etc. from the data.

A sample annotation guideline might cover the following content for our exercise:

1. Check for terms that might directly express emotions ("terrified", "happy", "sad” ...) or opinions ("good", "well", "awful", ...). If the majority of such words direct to one of the poles (positive or negative), choose that label (r15 is a good example for this). Keep in mind that sometimes expressed emotions and opinions might lead to a reverse meaning by negation (r1 is a good example for that).
2. Check for terms that imply comparisons ("prefer", "than", "in comparison to"). In the existence of such cases, the label can be decided if the relationship between the compared entity and the evaluated entity is explicitly revealed (r18 can be a partial example for this).
3. In general, classifying reviews with tags that refer to different parts of the movie content before moving on to the actual labeling section might be helpful. For example we can use tags such as "characters", "acting", "storyline" or "movie" to define what the review is actually aiming, and then try to analyze how the review evaluates each of these sub-elements. For instance r7 includes positive statements about both the acting performance and the movie as an end product. Having positivity about more than one of these sub-elements would lead to a strengthened evaluation decision (in this case, 'positive').
4. Due to the subjective nature of annotation, it might be a good idea to simply use hard-coded rules to restrict/allow labeling depending on certain criteria. Some examples might be:
   * If the review can only be tagged with a certain sub-element, avoid annotation. (As an example, imagine a review like "This movie is definitely intended for people who have watched Ice Age". It tries to give some general information about the movie by using the example of another movie, thus requiring a broader context. If the guidelines provide a certain tag for such situations, we can skip this review)
   * Potential sarcasm/irony should be supported by the presence of punctuation such as '[!]', '!?', '⸮'. Without punctuation support, annotators who feel that any given review is sarcastic/ironic, should mark the example without any label

In addition to the guideline, computing the inter-annotator agreement between the annotations of different human annotators is also widely applied in annotation tasks. The agreement can be calculated by using the percentage agreement (i.e., percentage of the agreed items over all the items). In the literature, there also exist different metrics such as Cohen's kappa and Krippendorf's alpha for measuring the agreement. There is an extensive overview of these metrics for the computational linguistics tasks [here](https://www.aclweb.org/anthology/J08-4004.pdf).

Increasing the number of annotation labels as in [E1b], will probably decrease the agreement score. You can prefer to use weighted agreements for this kind of labelling to differentiate the type of disagreements such as "positive vs somewhat positive" vs "positive vs negative". Methods to apply weighted differences into the agreement evaluation are also mentioned in Artstein&Poesio's paper referred above.

**[E2]:**

Since you don't speak Turkish, one way to classify the texts would be to try to recognize the words which look like the words you are familiar with. Names (Burak Yilmaz), acronyms (USB), international words (bank, futbol?, gol?) might be good candidates for indicating some familiarity.

These keywords might be helpful in classifying the texts. You can represent each document in terms of the count of these *familiar* words and compute the Naive Bayes probability ordering for the documents by assuming as P(gol | sports) > P(gol | technology) & P(gol | sports) > P(gol | entertainment), for instance.

**[E3]:**

In [E2] if you follow the keyword approach mentioned above, you do a Naive Bayes classification with a simplified set of features (i.e., keywords here) and you don't take the order of the words into consideration (i.e., bag of words). In this sense, the Naive Bayes computation is very similar to the n-gram modeling. We actually calculate the unigram probabilities conditioned on the output labels. And if we can only find a few words with high unigram probabilities conditioned on a given label, our model would lead to the correct label.

**[E4]:**

The impact of using highly correlated features (e.g., commonly co-occuring words) as independent features in Naive Bayes classification is nicely explained in:

<https://www.youtube.com/watch?v=_Niy2f-M9KA&feature=emb_title>

Video. Three words (Monaco, Hong, Kong) are used to classify whether the text is about Europe or Asia. Since *Hong* and *Kong* are always co-occurred in the texts, the impact of these words in the final classification score is duplicated. In order to avoid such complications, it might be a good strategy to combine the highly correlated features into one feature or to build a discriminative model such as using Logical Regression instead of Naive Bayes.

**[E5]:**

Please refer to your text book and Piazza threads (Question @20) for the solution of this exercise.

**[E6]:**

Please refer to your text book and Piazza threads (Question @20) for the solution of this exercise.