THE UNIVERSITY OF SYDNEY

ASSIGNMENT

ONE

Annotation task

Due date: 5pm Monday of week 5 (2018-04-09)

This assignment is worth 10% *of your final assessment.*

In this assignment, you will manually annotate news articles with a classification scheme aimed at the following scenario.

Our sample of articles comes from the Sharing News Online project, which tried to look at why people shared online news media on Facebook and Twitter. Here we have a sample of articles that were shared over a few months from various news web sites. (Note that the articles are copyright to their original authors and publishers, and we believe that the present is Fair Use.)

The researchers want to know: what topics of article get shared? Do these change over place? Over time? Does topic help differentiate between ”been shared” and ”went viral”? In order to do that analysis, we need to first classify each article in their collection according to its topic.

The assignment runs in multiple phases:

Throughout	Think and write about decisions made and challenges
Week 2	Discuss/design classification scheme in labs and on Canvas
Tuesday week 3	Classification scheme released; begin annotating
Monday week 4	Common annotations due
Tuesday week 4	Common annotations released for evaluation; begin evaluating agreement
Monday week 5	Report due

The classification scheme is drafted by the class in the second lab session. The third lab session includes a training/discussion session for the annotation task. Each student then annotates 160 entries:

- 80 that everyone will annotate
- the remaining articles will each be annotated by 3 students, assigned randomly

The first 80 articles will be used for your analysis of inter-annotator agreement in this assignment. All annotations will be used as training and evaluation data for the second assignment, yielding $(80 + (160 - 80) \times \#\text{students} \div 3)$ distinct training examples in total.

This structure helps us assure quality and evaluate how well annotators can agree on this task. You do not need to worry about it too much: just annotate your first 80 entries by Monday week 4, and the remainder by the time the assignment is due.

It should not take more than 30 seconds on average to classify each article, so this should not take you very long. We recommend you complete it early so you can focus on the report.

However, the assessment is not about you *doing* the annotation. It is about your understanding, analysis and critique of the task, annotation scheme, annotation process and user interface, and your inter-annotator agreement on the 80 common articles annotated by every student in the class.

You will be assessed on a short report that discusses the annotation task both quantitatively and qualitatively. The **quantitative analysis** should use *inter-annotator agreement* statistics to measure your agreement with the rest of the class and to estimate human performance on the task. The **qualitative analysis** should present a critique of the annotation task, the annotation scheme, the annotation process and web-based annotation tool, and the problem more generally of assigning categories to complex linguistic and ontological phenomena.

1.1 Report structure

Your report should have the following sections:

Results For each metric listed in section 1.4 (confusion matrix, class-wise precision, class-wise recall, class-wise F-score, averaged P or R or F, Cohen's kappa, Fleiss' kappa):

- Describe what the metric is useful for;
- Show your results, making sure any tables are clearly labelled; and
- Provide some brief quantitative analysis: remark on something that the statistic tells you about the annotated dataset. This may be something the statistic tells you alone (relative to how high you should expect it to be for good quality), or something the statistic tells you when compared to another statistic.

Discussion Make one reflective/critical argument about the process of building a classification dataset under each of the headings {*Annotation task*, *Annotation scheme*, *Annotation tool*, *Discussion and agreement*} (see section 1.5 for ideas):

- Please precede each argument with a question you are responding to. It can be one of the questions from section 1.5 below, or could be your own question.
- Where appropriate, each argument should be supported by quantitative evidence (i.e. refer back to Results) or examples (i.e. mention a specific example by its title, and the kinds of issues it raised).
- At least one argument should refer to a specific example from the dataset. At least one should refer to a quantitative result.

Overall the report should take about three pages, excluding tables and plots.

1.2 Submission instructions

Your submission will consist of two components:

1. **Annotation**: To be completed in a Google Spreadsheet you will be invited to edit before the Week 3 lab. The annotation scheme (the set of categories being applied) will be published within the spreadsheet and on the course site. Please complete annotating the set of 80 common articles **by the end of Monday Week 4**.
2. **Report**: To be submitted as a PDF or Microsoft Word document to Turnitin through Canvas Assignments. *Your submission is not complete until you receive a "digital receipt".*

No work should be submitted after the deadline without explicit prior arrangement. Please see the course outline and slides from the first lecture with respect to late submission and special consideration policies.

1.3 Annotation task

The task is to annotate articles according to a specified classification scheme. The task is to assign only a **single category** to each article, even when there is ambiguity, or when multiple categories may seem more appropriate. In these cases, you should choose the category that is most prominent or useful for our application (or use an "other" category if we choose to have one in our schema).

The annotation tool will be a Google Spreadsheet that you will all be shared into. Each student will have a separate sheet in the spreadsheet. Please ensure you only modify cells in your own sheet. The version tracking feature in Google Spreadsheet should protect against accidental mistakes.

Each student in COMP5046 must annotate their 160 articles as follows:

- 80 annotated by every student
- the remainder each assigned to 3–4 randomly selected students

This ensures that we have multiple annotations per article for reliability purposes, but also that we'll have a large pool of distinct annotated articles in total. The annotations for the first 80 articles will be used for your analysis of inter-annotator agreement (see below).

Any articles that are not correctly extracted – they aren't in English, or they don't have substantial text content within the provided snippet – should be marked with category *Error*.

1.4 Evaluation and results: quantitative analysis

There is no absolute *correct* answer for annotation tasks, only the best answer as agreed between the annotators. Usually agreement can be reached after discussing the options and a new policy is documented for similar future cases – like we did for some of those cases in the second lab. However, sometimes the annotators just have different intuition about the case and so agreement cannot be reached.

For our purposes, we will assume that the best answer is the one chosen by most annotators (i.e. each annotator gets one vote). Therefore the first step in the quantitative analysis is to decide what is the most frequent category for each article.

Ties are very unlikely with so many annotators as we have here. In case of a tie, choose a tying category at random.

We'll refer to the result of this process as the *voted* annotation. There are four different inter-annotator agreement statistics that you need to calculate as part of the quantitative analysis. For most of these statistics, you will compare your annotations against the voted annotation, while for Fleiss' Kappa, you compare all annotators against each other.

You are only calculating these statistics over the first 80 articles that everyone in the class has annotated.

1.4.1 Confusion matrix

A confusion matrix is a table that compares the voted choices with your choices over all of the articles. Each cell in the table records the number of times you classified an article into a given category compared to the voted category.

The confusion matrix Wikipedia article gives a simple example `http://en.wikipedia.org/wiki/Confusion_matrix`. Since confusion matrices are often used for evaluating the output of supervised machine learning systems, many explanations of confusion matrices will call the voted category the *correct, actual or gold* category; and will call your category the *predicted* category.

The cells on the diagonal of a confusion matrix are where the voted choice matches your choice. In a perfect annotation, all of the counts appear on the diagonal. The non-diagonal cells are where your choice does not match the voted choice, i.e. there was confusion between the classes. The larger the count in a non-diagonal cell, the harder those two categories are to distinguish.

Despite the availability of more summarised statistics described below, the confusion matrix contains a lot of information that is easy to interpret about the properties of the dataset and the kinds of disagreements present.

You need to create a confusion matrix where the *rows are the voted categories* and the *columns are your categories*. Please label which axis is which to help the reader.

1.4.2 Precision, Recall and F-score

Precision, recall and F-score are standard evaluation metrics for information retrieval and extraction systems. *Precision* is the proportion of articles you give a particular category that match the voted category (i.e. the proportion that are correct). *Recall* is the proportion of articles for a voted category that you actually annotate with that category. *F-score* is the harmonic mean of precision and recall. More information is available on the Wikipedia page: `http://en.wikipedia.org/wiki/F1_score`.

Let's consider a (fake) example with the category `Health`. Say 20 articles were classified as `Health` by you, and 25 articles were classified as `Health` by voting. If 18 articles are in common between the two sets (i.e. you agree with the voted category 18 out of 20 times), then precision is $P = \frac{18}{20} = 90\%$; the recall is $R = \frac{18}{25} = 72\%$ and the F-score is:

$$
\begin{aligned}
F &= \frac{1}{\frac{0.5}{P} + \frac{0.5}{R}} && (1.1)\\
&= \frac{2PR}{P+R} && (1.2)\\
&= \frac{2 \times 90\% \times 72\%}{90\% + 72\%} && (1.3)\\
&= 80\% && (1.4)
\end{aligned}
$$

For the report, you must calculate precision, recall and F-score *for each of the classes*, and an overall precision, recall and F-score. For overall scores you may report a micro-average (potentially excluding some "negative" class; otherwise it is identical to *accuracy* for multiclass evaluation), a macro-average, or a weighted average, as long as you show an understanding of how it summarises the data.

1.4.3 Cohen's Kappa

Cohen's Kappa (κ) statistic is a measure of inter-annotator agreement between two annotators over two or more categories. It was designed by Jacob Cohen (in 1960) to take into account the probability of randomly choosing classes that agree with the other annotator. For instance, if there are two equal probability categories, and you only reach 50% agreement with the other annotator, this is no better than if you both randomly annotated.

The formula for Cohen's Kappa is:

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)} \tag{1.5}$$

where $P(a)$ is observed probability of agreement (the proportion of the time you actually agree with the voted category), and the $P(e)$ is the expected probability of agreement. The expected probability is calculated using your and the voted probability distributions over the categories.

The Wikipedia article on Cohen's Kappa gives a very simple example for calculating κ that you should work through first: `http://en.wikipedia.org/wiki/Cohen's_kappa`.

1.4.4 Fleiss' Kappa

The final statistic you should calculate is Fleiss' Kappa statistic, which is a version of κ designed for multiple annotators.[1] Fleiss' kappa is best calculated by producing a table like the example on the Wikipedia page: `http://en.wikipedia.org/wiki/Fleiss'_kappa`.

Unlike the previous statistics, Fleiss' kappa will *not* compare your classifications to the majority vote, but will instead use everyone's individual classification decisions. To simplify the calculation, you can remove any annotator who has not completed all 80 annotations from the data.

Your report does *not* need to include the whole table of computed counts for calculating Fleiss' kappa, but should include your calculated value and enough of a description to explain how it works.

1.4.5 Data file

You will be provided a UTF-8 encoded comma separated value (CSV) file called **`annotations.csv`** on Tuesday in week 4. It will be available alongside the assignment description in Canvas. The **`annotations.csv`** file contains all of the classification decisions up to Tuesday Week 4. The file has three columns:

- the username of the annotator who made the decision.
- article ID (may contain non-ASCII characters); and
- the category chosen.

Please use a standard CSV reading tool to ensure fields containing commas or quotation marks are correctly parsed.

1.4.6 Code

You do not need to submit your code, and you may use existing libraries to calculate metrics.

1.4.7 Significant figures

Note: 80 is a small number. Do not report evaluation metrics to 9 decimal places, as they are meaningless.

[1] in fact, it's a generalisation of Scott's pi statistic

1.5 Discussion: qualitative analysis

The purpose of this assessment task is for you to experience the complexity of creating an annotation task, defining the annotation scheme and deciding between edge cases that are hard to annotate.

In the qualitative analysis, you should talk about the problems you and your colleagues have experienced and discussed as part of the annotation process. Some of the things you might like to discuss include:

Annotation task Does it make sense to annotate a single category per article? (What are the advantages and disadvantages in terms of annotator accuracy and efficiency, building machine learning systems to use the classification, and end-applications that use the classification output? These kinds of questions can be asked of all task/scheme design decisions, such as those listed in Week 2's lab.)

Annotation scheme Were these the correct categories to use? Were there important categories missing? Were any of the categories too broad, too narrow, or completely redundant? Which categories were difficult to distinguish between? Was a flat scheme the right decision, or should the categories be hierarchical?

Annotation tool Did the tool make annotation efficient and accurate? Are there ways you would improve the tool to make it faster and less error prone to use?

Discussion and agreement Did you find the discussions helpful? Were you surprised by some of the category decisions of other students and/or how well/poorly you agreed with everyone else?

Finally, this task might not be particularly difficult, and was completed by well-educated annotators. What implications does the difficulty of annotation have on building systems that rely on human annotation?

The most important thing we are looking for in this assessment task is *sophistication of understanding*, rather than the amount of discussion. You should use examples as much as possible to support your discussion, and ideally draw conclusions across both the quantitative and qualitative analysis.