# HW-MIDTERM

Sam Tokheim

May 1, 2013

## 4 Normalize the following feature values

Use min-max [0-1] normalization, output fractions or 4 decimal places.

**[5, 25, 6, 12, 15]**

**--**

For the midterm, I missed the “\* (D-C)+C” part of the formula and I incorrectly put the (max(dataset)-value) rather than the (max(dataset)-min(dataset)) in the denominator which resulted in the following incorrect output: [0/20, 20/0, 1/19, 7/13, 10/10]. Here is the correct formula and output.

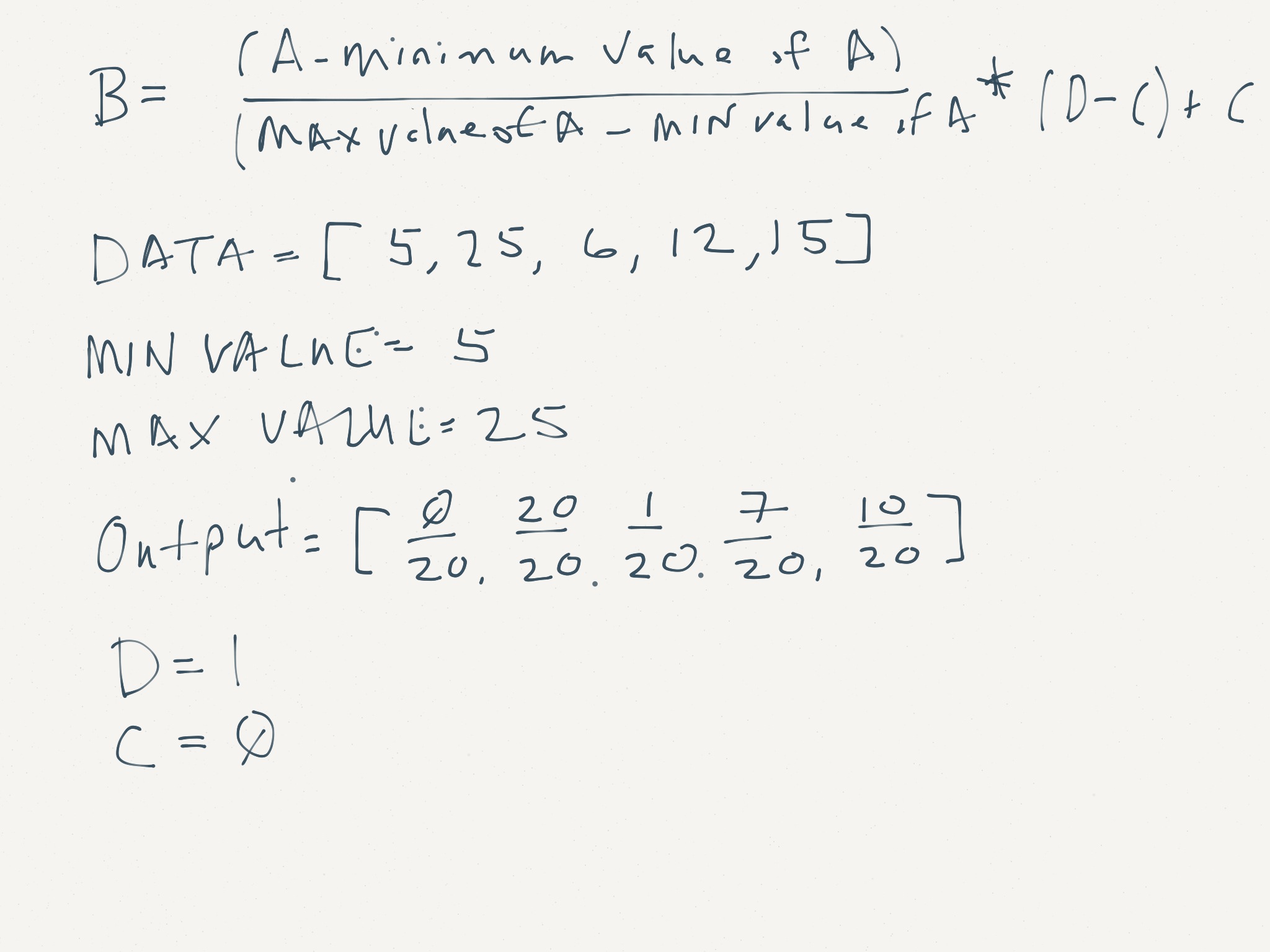
****

Figure : I used this blog for help with this the min-max formula (U, Venkatesh)

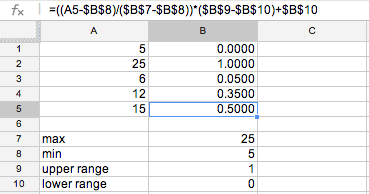


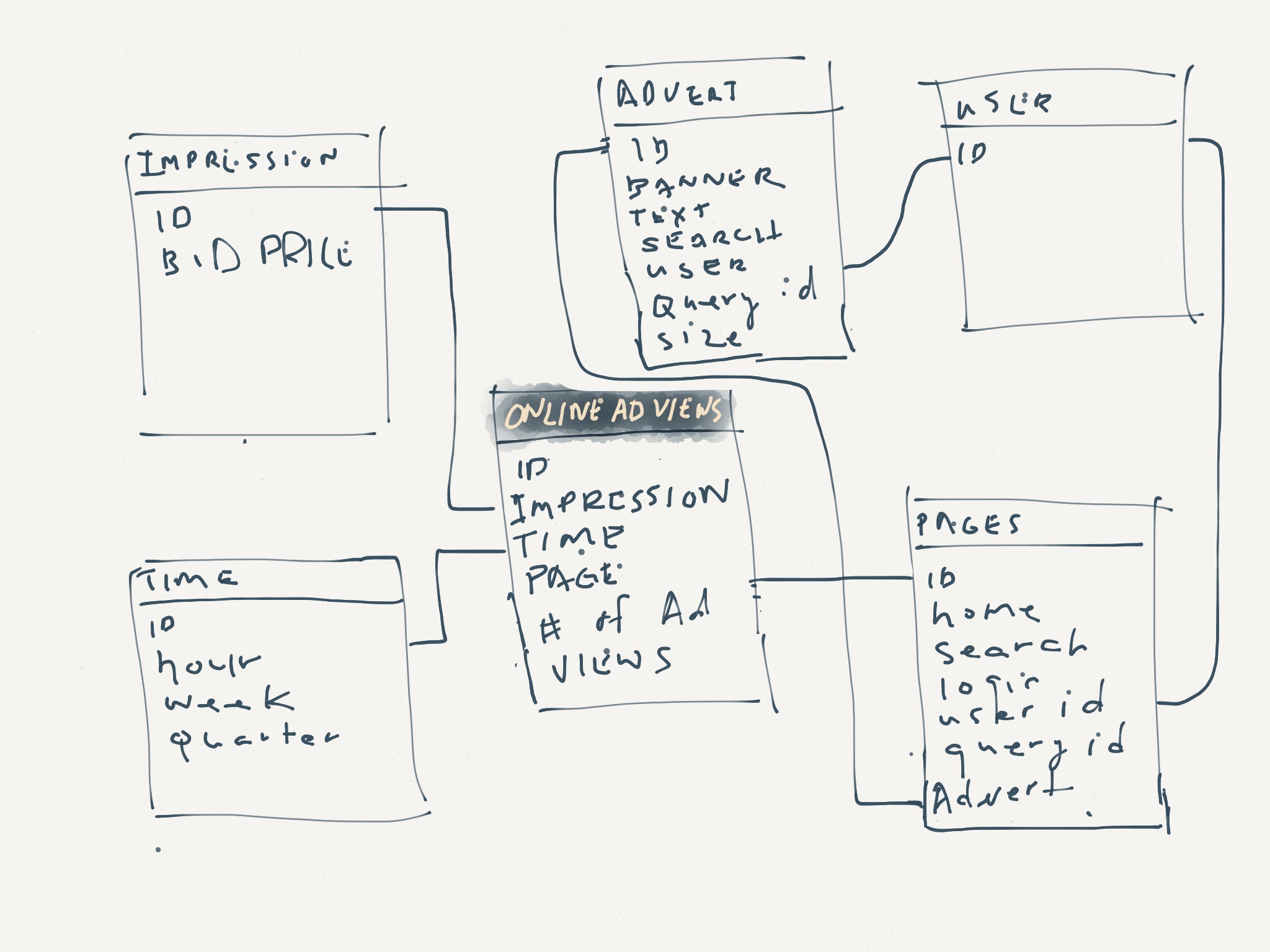
Figure : Excel version

## 6 Draw a star schema for the following information needs:

* The fact table tracks online advertising views
* Advertisements can be different types. Types have names (e.g. home, search, or login) and IDs (e.g. the user id, query id)
* The highest time resolution we need is 1 hour, but we want the ability to summarize by week and quarter
* Every advertisement view is also associated with bid price paid for the impression.

--

On the exam, I neglected to add a table that tracks the ‘time resolution’. I think that I was also confused by how to link the keys between ‘Online Ad Views’, ‘Advertisements’ and ‘Pages’. I think that the correct way to link these is illustrated in my updated Star Schema below where Pages have a key to Advertisements *and* Online Ad Views. And Advertisements do not link to Online Add Views.



## 9 Describe precision and recall in English

On the midterm, I describe precision but did not adequately describe recall. Here’s what I wrote about precision:

“A perfect precision score of 1 for a class means that the classifier labeled as belonging to class C does, indeed, belong to that class. There is a tension between precision and recall. Generally speaking, with greater precision comes less recall and vice versa”.

Using the example of a Yelp search to illustrate, recall can be defined as ‘the number of relevant documents retrieved by a search divided by the total number of existing relevant documents (Wikipedia contributors, 2013)’ and precision can be defined as ‘the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search’ (Wikipedia contributors, 2013). It would be easy to have a recall score of 1.0 if, when a user searched for ‘sushi’, Yelp simply returned all restaurants (not just Sushi restaurants). Although, all of the sushi restaurants would be included in the search results, a bunch of non-relevant restaurants would also be included. Therefore, it’s important to consider the precision of the yelp search results not just the recall.

## 10 What is information gain and how is it used to build decision trees?

Information gain is a measure for deciding the relevance of an attribute in a decision tree. It compares the entropy (or how mixed) results are before and after a split.

**“A decision tree can be constructed top-down using the information gain in the following way** (Wikipedia contributors, 2013)**:**

1. Begin at the root node
2. Determine the attribute with the highest information gain which is not already used as an ancestor node
3. Add a child node for each possible value of that attribute
4. Attach all examples to the child node where the attribute values of the examples are identical to the attribute value attached to the node
5. If all examples attached to the child node can be classified uniquely add that classification to that node and mark it as leaf node
6. Go back to step two if there are unused attributes left, otherwise add the classification of most of the examples attached to the child node

(Wikipedia contributors, 2013)”

**12 When implementing a neural network trainer, what are our terminating conditions?**

What are (at least) three ways we can choose to stop iterating?

--

On the midterm, two of my terminating conditions were too similar and so, I only had 2 of the 3 required answers. Here’s what I wrote:

1. We can simply choose to stop iterating after arbitrarily choosing n iterations
2. We could choose to stop running when the error of the output layer reaches a tolerance that we specify.

Here’s the third terminating condition:

1. We could terminate when the changes in weights is small. This would be a number that we would define.

**14 Draw a dendrogram showing clustering for these points**

Points are agglomeratively clustered using L2 norm, minimum cluster distance.

--

My understanding is that we start by connecting the closest nodes and then repeat until all of the nodes are connected as illustrated in the second dendrogram image below.



Figure : Points

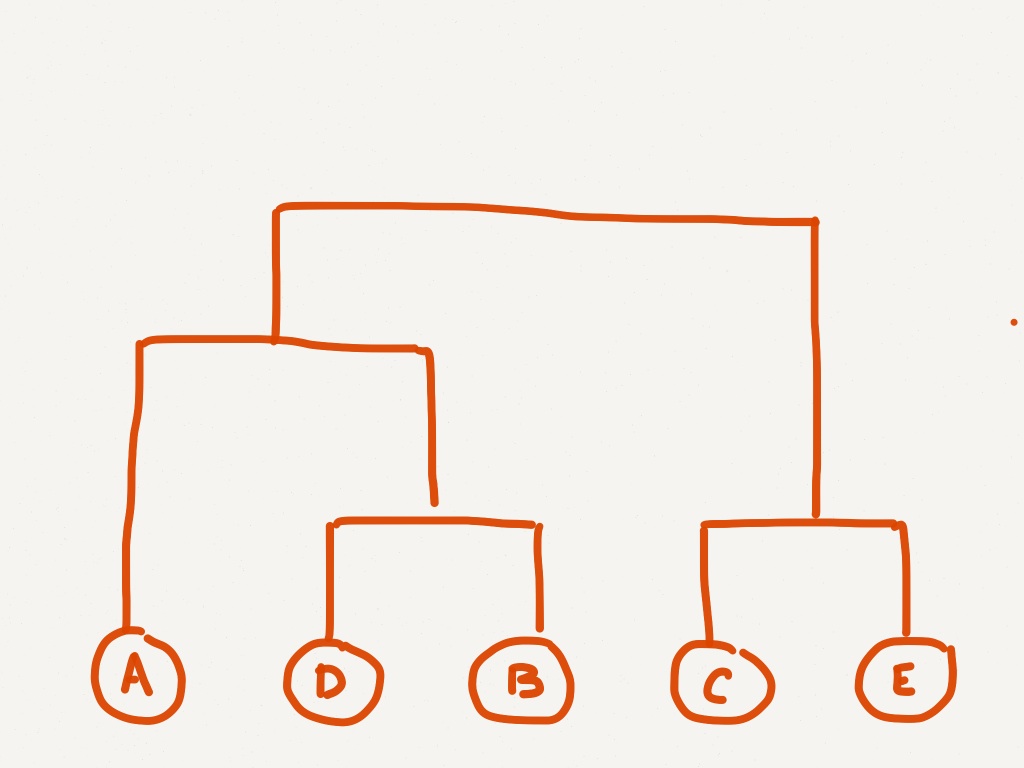


Figure : Dendrogram showing clustering for points

**15 Write MapReduce inputs/outputs**

We’d like to use the MapReduce paradigm to answer the question: What was the average review star rating for a business category? You’re using the business record type, which has

* Average number of stars
* List of categories
* Number of reviews

Write the key/value signatures for the MapReduce steps you’d write (you may optionally explain what each step does in English or pseudocode. Key/Value signatures are:

Map | Reduce step\_name: type of key, type\_of\_value => type\_of\_key, type\_of\_value

For example, in a word count job, part of the solution might be:

Map get\_words: None, record => review\_id, number of words

--

For this question, I created an actual MapReduce job and pushed the code and results to my Git repository for this class. Here is a screen shot of my code:

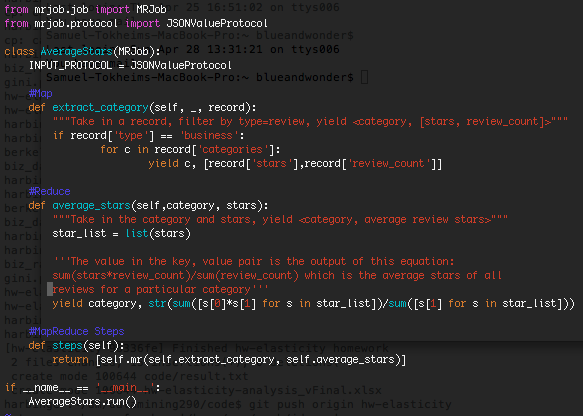


Figure : MapReduce code

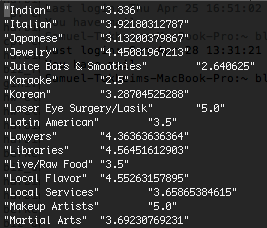


Figure : Sample of MapReduce output using a subset of the Yelp dataset.

**16 What is the Jaccard index between these two sets?**

* {good luck on the midterm}
* {the midterm is over}

**--**

The Jaccard index compares the ‘similarity and diversity of sample sets (Wikipedia contributors)’ and is defined as ‘the size of the intersection divided by the size of the union of the sample sets (Wikipedia contributors)’.

So…

**Intersection =** words incommon

**Union =** words in either

**Words in common (Intersection)** = the, midterm

**Words in either (Union)** = good, luck, on, the, midterm, is, over

Size of intersection = 2

Size of union = 7

**Jaccard index =** Size of Intersection / Size of Union = 2/7

# Works Cited

U, Venkatesh. (n.d.). *Data Preprocessing – Normalization*. Retrieved 5 2, 2013, from Data Mining Blog: http://intelligencemining.blogspot.com/2009/07/data-preprocessing-normalization.html

Wikipedia contributors. (2013, 4 19). *Information gain in decision trees*. Retrieved 5 2, 2013, from Wikipedia, The Free Encyclopedia.: http://en.wikipedia.org/w/index.php?title=Information\_gain\_in\_decision\_trees&oldid=551143906

Wikipedia contributors. (n.d.). *Jaccard index*. Retrieved 5 2, 2013, from Wikipedia, The Free Encyclopedia.: http://en.wikipedia.org/w/index.php?title=Jaccard\_index&oldid=549575590

Wikipedia contributors. (2013, 3 18). *Precision and recall*. Retrieved 5 1, 2013, from Wikipedia, The Free Encyclopedia.: http://en.wikipedia.org/w/index.php?title=Precision\_and\_recall&oldid=545239298