Lab 4: Map/Reduce

Import the data

For this lab we will need the data base *twitter* with the collection *tweets*. This collection contains a sample of 53641 tweets.

1. Import the data base from the shell as following:

```
> mongoimport --db twitter --collection tweets --file twitter.json
```

2. If the import was successfull the data base and collection have been created and filled automatically. Please use a Mongo Shell to verify the correct import of the data:

```
> db.tweets.count()
53641
```

Exercise: Twitter Trends with Map/Reduce

Twitter users use hash tags (meaning #) in their texts to mark keywords, e.g. #justinbieber. Using these hash tags is a good way of determining what's currently hot on Twitter and actually Twitter is doing exactly this for determining the current Twitter Trends.

Now we want to do exactly the same but with MongoDB and Map/Reduce.

In doing so, we first need a map() function, which splits the text of a tweet and counts all the hash tags # occurring in that text. The body of the method will be something like this and you need to fill out the logic:

```
// Map function
map = function() {
      if (!this.text) {
            return;
      }
```

```
        // Determine the hash tags

        …

        // Count each hash tag and emit it to the reduce function
        emit(hashtag, 1)
};
```

Next we need the reduce() function, which sums up all the hash tags found by the map() function. The body of the function could look like this and again it is your job to create the logic of the function:

```
// Reduce function
reduce = function(key, values) {
    // Sum up the hash tags
      …
};
```

Now we have what we need and start Map/Reduce to determine the trend on this Twitter collection. We will store the results of Map/Reduce in a collection named *twitter_trends*:

```
> db.tweets.mapReduce(map, reduce, { out : "twitter_trends"} );
```

After the successful execution of our Map/Reduce function we can determine the Top-10 of the most used hash tags in this collection as following:

```
> db.twitter_trends.find().sort({"value" : -1}).limit(10);
```

Question: What is the most used hash tag in this Twitter collection?