

LINEE GUIDA SCHEMA MONGO DB

La Meo quest'anno ha specificato delle collezioni consigliate per MongoDB, cosa che anno scorso non era previsto! Per cui presumo che le abbia messe apposta poiché non soddisfatta delle soluzioni adottate dagli studenti negli anni passati. Di seguito, ti descrivo nel dettaglio l'interpretazione del contenuto di queste risorse.

- Creerete una collezione per memorizzare il contenuto dei microblogs di Twitter.
Twitter:

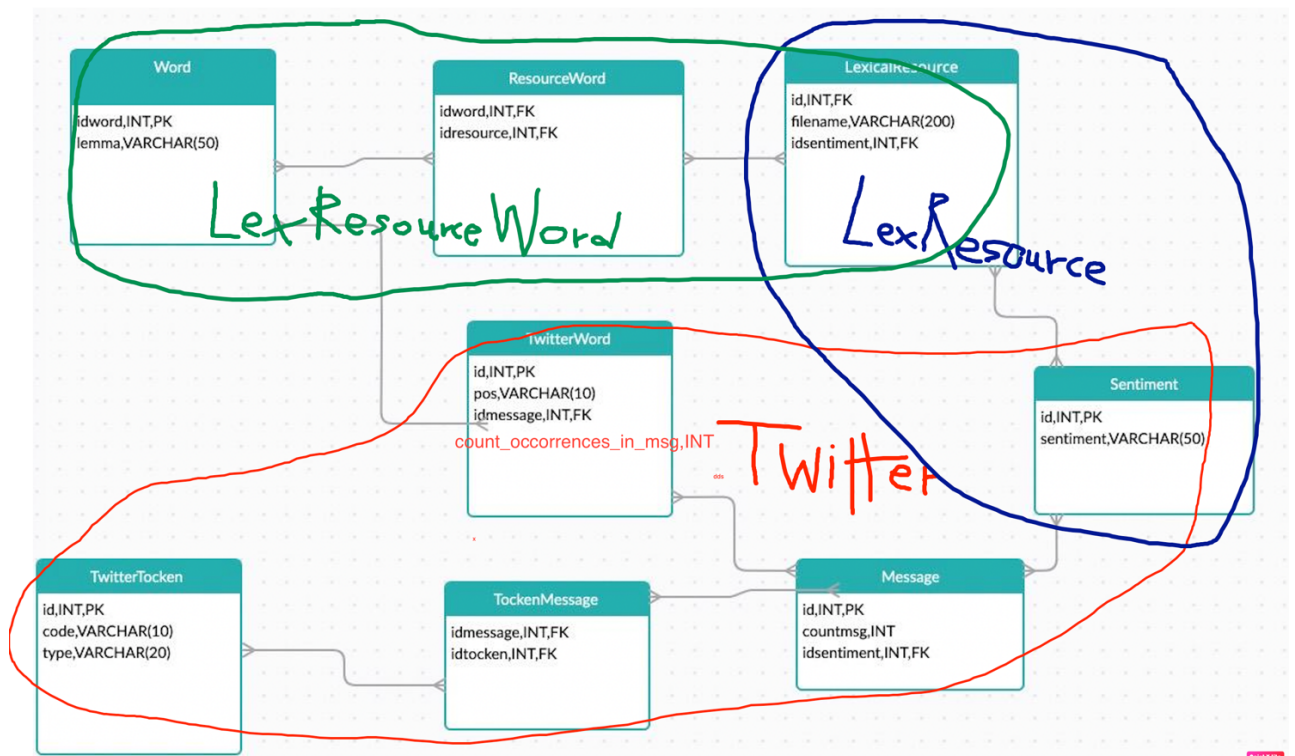
```
{
  sentiment : ["joy","sadness","trust","anger","anticipation","disgust","fear","surprise"],
  [[doc_number : order_number_of_the_document_in_the_sentiment_file,
  words : [{lemma : "the_lemma", POS : posTag, freq : count_occurrences_in_msg,
            in_lex_resources : {$ref : LexResourcesWords, $id : reference_Object_id}
            },
            ...]
  ],
  hashtags : [list_of_hashtags],
  emojis : [list_of_emojis],
  emoticons : [list_of_emoticons]],
  ...
}]
```

Fanno riferimento alla collezione seguente
- Collezione delle parole nelle risorse:
LexResourcesWords

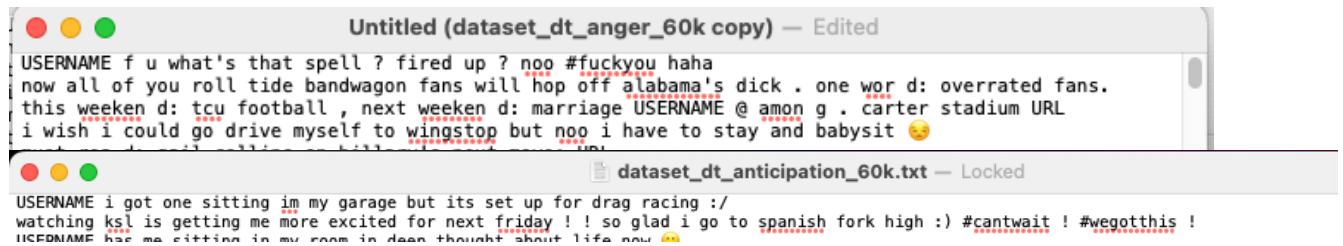
```
{_id: ObjectId("..."),
 lemma : word,
 resources : [ {$ref : LexResources, $id : resource_filename}, {},... ]
}
```
- Nome delle risorse e numero totale di parole:
LexResources

```
{_id : resource_filename,
 sentiment: nome_sentimento,
 totNumberWords : number_of_words}
```

Relazione tra le 3 collezioni proposte (**Twitter**, **LexResourcesWord**, **LexResources**) con lo schema MySQL attualmente implementato



Messaggi Twitter



Risorse Lessicali



Twitter

```
{sentiment: "anger",
 doc_number: 1,
 words : [ {lemma: "spell", POS: Noun, in_lex_resources: [ reference_Object_id = 3 ]},
 {lemma: "fire", POS: Verb, in_lex_resources: [ reference_Object_id = 5 ] }, ... etc]
 hashtags: [ "#fuckyou" ]
}

{ sentiment: "anger",
 doc_number: 2,
 words : [{lemma: tide, POS Noun}, {lemma: fun, POS Noun, freq :2, in_lex_resources: [
reference_Object_id = 7 ] }, ], ... etc]
 emoticons: [ "d:" ]
}

....

{sentiment: "anticipation",
 doc_number: 1,
 words : [ {lemma: "garage", POS: Noun}, {lemma: "racing", POS: Verb}, ... etc]
 emoticons: [ ":/"]
}

{ sentiment: "anticipation",
 doc_number: 2,
 words : [{lemma: excited, POS: Adjective}, ... etc]
 emoticons: [ ":", ""]
 hashtags: [ "#cantwait", "#wegotthis" ]
}
```

LexResourcesWords

```
{
  _id: ObjectId("...") = 1
  lemma: "abundance",
  resources: ["NRC_anticipation.txt", .... etc]
}
{
  _id: ObjectId("...") = 2
  lemma: "accelerate",
  resources: ["NRC_anticipation.txt", .... etc]
}
{
  _id: ObjectId("...") = 3
  lemma: "spell",
  resources: ["NRC_anticipation.txt", .... etc]
}
{
  _id: ObjectId("...") = 4
  lemma: "aloe",
  resources: ["NRC_anticipation.txt", .... etc]
}
{
  _id: ObjectId("...") = 5
  lemma: "Fire",
  resources: ["NRC_anticipation.txt", .... etc]
}
{
  _id: ObjectId("...") = 6
  lemma: "amour",
  resources: ["NRC_anticipation.txt", .... etc]
}
{
  _id: ObjectId("...") = 7
  lemma: "Fun",
  resources: ["NRC_anticipation.txt", .... etc]
}
```

LexResources

```
{
  _id: "NRC_anticipation.txt",
  sentiment: "anticipation",
  totNumberOfWors: 839
}
```

Note:

- Essendo *MongoDB schemaless*, se in un particolare messaggio non sono presenti ad esempio emoticons, non bisogna aggiungere l'attributo (vedi messaggio1 in cui non è stato inserito l'attributo emoticons)
- il campo **freq** indica il numero di volte che uno stesso lemma si ripete all'interno dello stesso messaggio Twitter. Dobbiamo decidere se eventualmente sia utile inserirlo o meno. Ad ogni modo, dato che MongoDB è schemaless, possiamo inserirlo solo per quei messaggi (quindi records) in cui la parola si ripete più volte.

ES:

I play the piano but I also play the guitar.

In questo caso il lemma play all'interno dello stesso messaggio si ripete più volte.

- Il campo totNumberWords in MongoDB è ridondante (infatti non è stato definito nel modello relazionale). Tuttavia, la Meo lo ha definito perché è possibile farlo nei modelli nosql, allo scopo di migliorare eventualmente le performances delle queries.
- ObjectId dovrebbe essere una specie di autoincrement, utilizzabile per assegnare un valore ad un record di LexResourceWords: <https://www.geeksforgeeks.org/what-is-objectid-in-mongodb/>