Uppsala University

Cloud computing

Cloud computing - Count pronouns in tweets

Authors:

Alessandro Piccolo

October 8, 2016

UPPSALA
UNIVERSITET

1 Introduction

A prototype system was built to analyze a dataset of Twitter tweets collected beforehand using Twitter's datastream API. The tweets were available in the public container 'tweets' in the SSC cloud, and the dataset consisted of a number of files containing lineseparated tweet entries. Each tweet is a JSON document. The main task was to count and visualize the Swedish pronouns "han", "hon", "den", "det", "denna", "denne" and the gender neutral, new pronoun "hen". The solution was based on the distributed task queue 'Celery', using 'RabbitMQ' as the broker and Flask to visualize and request tasks. The basic model is shown in figure 1

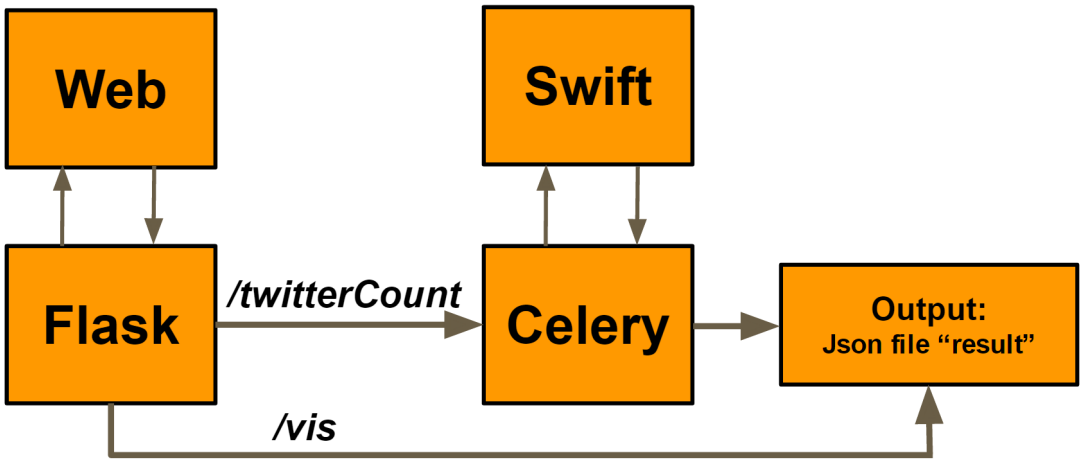

Figure 1: Model of the system

2 Results

Figure 2 shows the started flask server and one can see that it has been pinged once by figure 3. Figure 4 shows how the celery worker is actually working. The data is visualized as showed in figure 5.

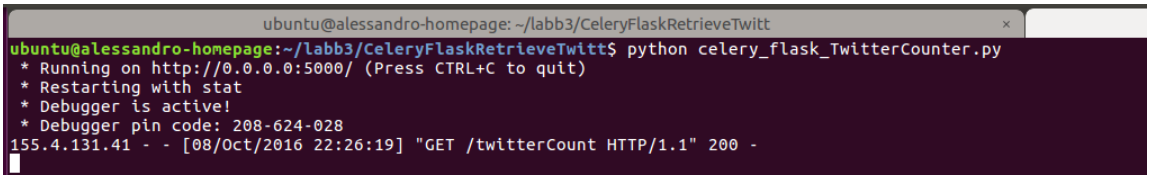

Figure 2: Flask server

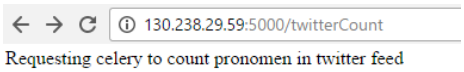

Figure 3: User request for counting pronouns to celery worker

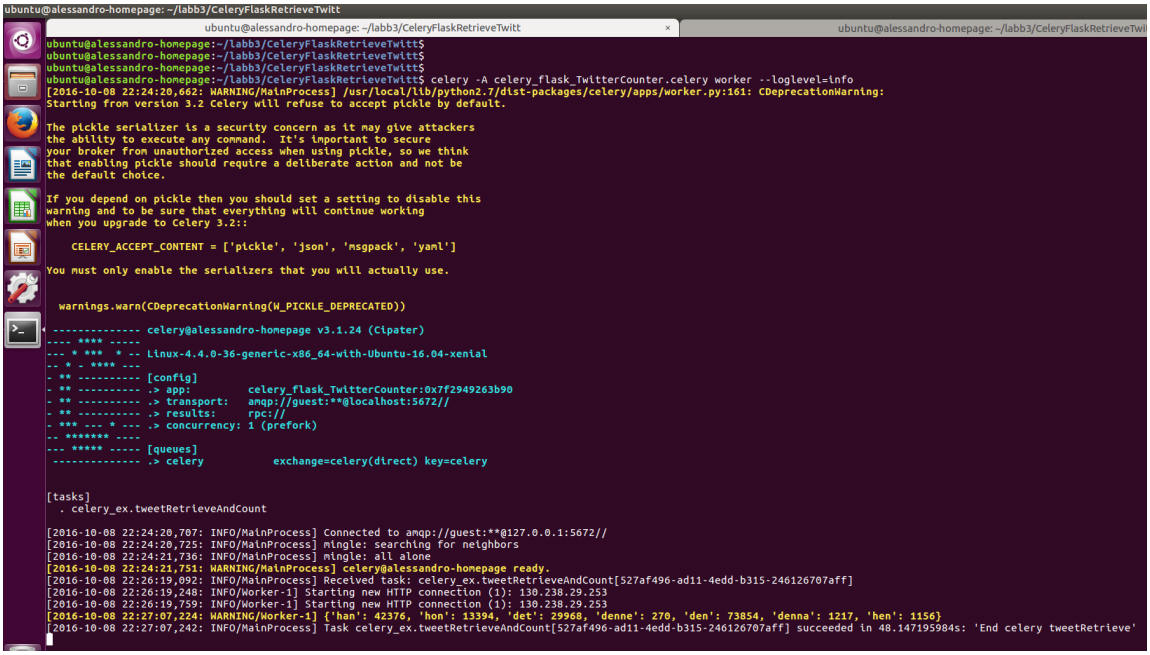

Figure 4: Visualization of celery worker in action

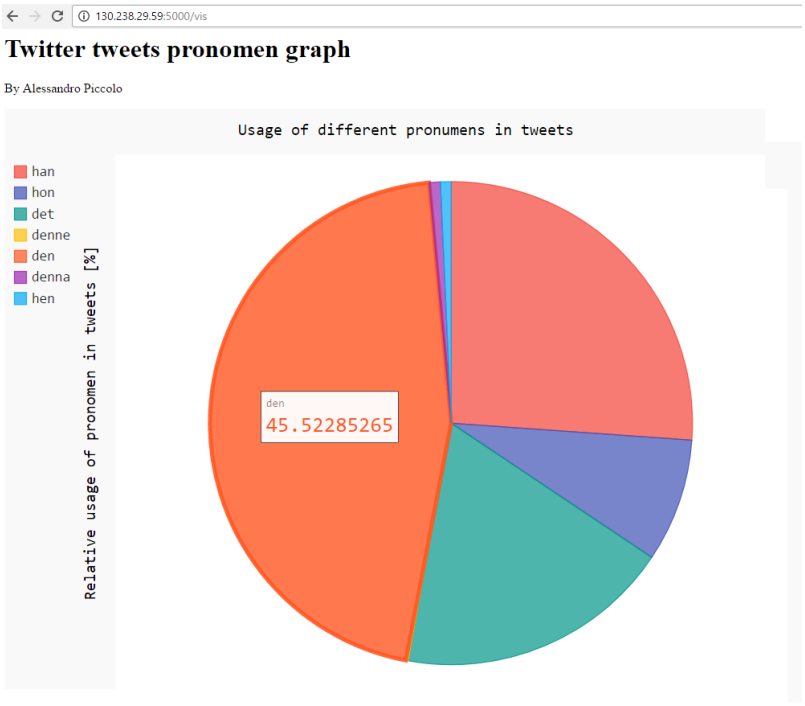

Figure 5: Visualization of pronouns count with flask pygal packet

3 Methods of celery_flask_twitterCount.py

This python file creates a worker and a flask server. It has two methods for the flask server, twitterCount() for requesting to start counting the pronouns for a celery worker and vis() for visualizing the count. It uses basic html with pygval package to visualize the json file called results (if there is no result file then it will tell the user to request and start a count). It also has a celery method called tweetRetrieveAndCount() which counts the use of pronouns in unique tweets in the files of the swift container in the cloud.

4 Packets to install

Basically these are the following commands that needs to be executed for a complete installation.

$ sudo apt-get update
$ sudo apt-get upgrade
$ sudo apt-get install rabbitmq-server
$ sudo locale-gen sv_SE.UTF-8
$ sudo apt-get install python-pip
$ sudo pip install celery
$ sudo apt install python-celery-common

$ sudo pip install flask
$ sudo pip install python-swiftclient
$ sudo apt-get install python-keystoneclient
$ sudo pip install pygal

5 How to use

The python files can be found in *https* : *//github.com/AlessandroPiccolo/labb*3*.git*
under labb3/CeleryFlaskRetrieveTwitt/.

Basically, open 2 terminals and ssh to instance, source g2015034-openrc.sh in each terminal.
Terminal 1: celery -A celery_flask_TwitterCounter.celery worker –loglevel=info
Terminal 2: python celery_flask_TwitterCounter.py

The user can initiate the request to count pronouns by accessing the following URL.

- http://floatingip:port/twitterCount

The previous request produces a json file called result. If there is such a file then one can visualize it by accessing the following URL.

- http://floatingip:port/vis

Appendices

A Github repository

The following files can be found in the following git repository,

https : *//github.com/AlessandroPiccolo/labb*3*/tree/master/CeleryFlaskRetrieveTwitt*