The University Of Sheffield.

Trend Miner

25.09.2012

UNIVERSITY OF Southampton

Trendminer: An Architecture for Real Time Analysis of Social Media Text

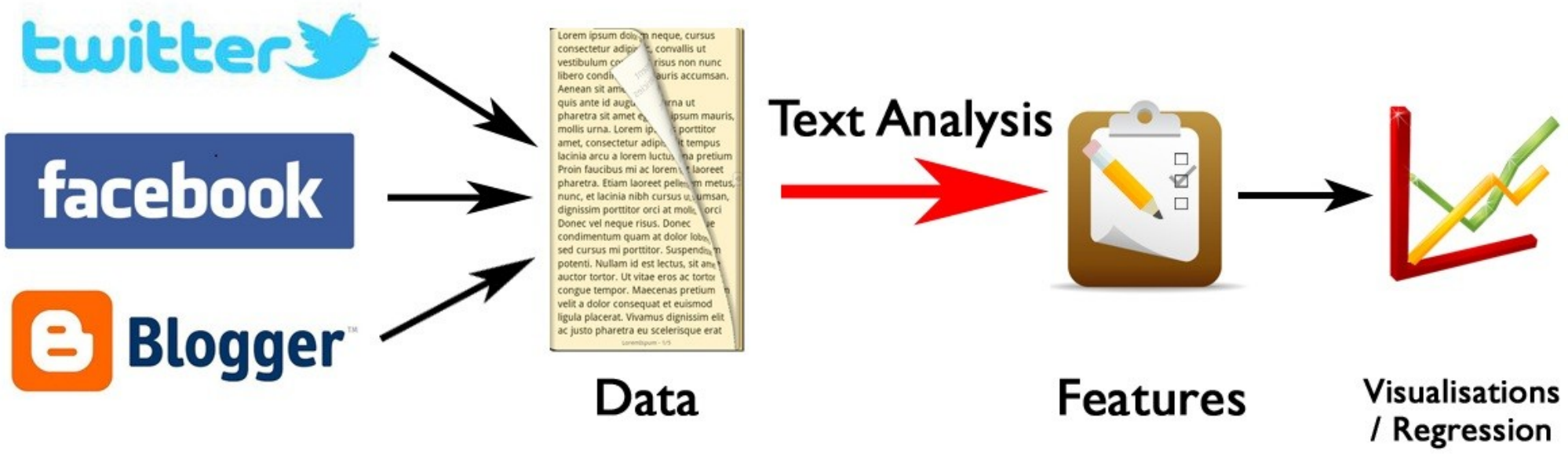
Daniel Preoțiuc-Pietro, Sina Samangooei

Trevor Cohn, Nicholas Gibbins, Mahesan Niranjan

Motivating Example

RT @MediaScotland greeeat!!! lvly speech by cameron on scott's indy :) #indyref

Background



Texts are short and different in style than from traditional sources

Real Time Architecture for Text Processing

We aim to integrate existing and new tools for OSN data processing in a framework that is:

Fast – real time processing

Modular - easy to add/change modules

Pipeline architecture - flexible to the user's needs

Extensible - different sources of data (e.g. Facebook)

Architecture

I/O bound: analysis takes less than random disk access

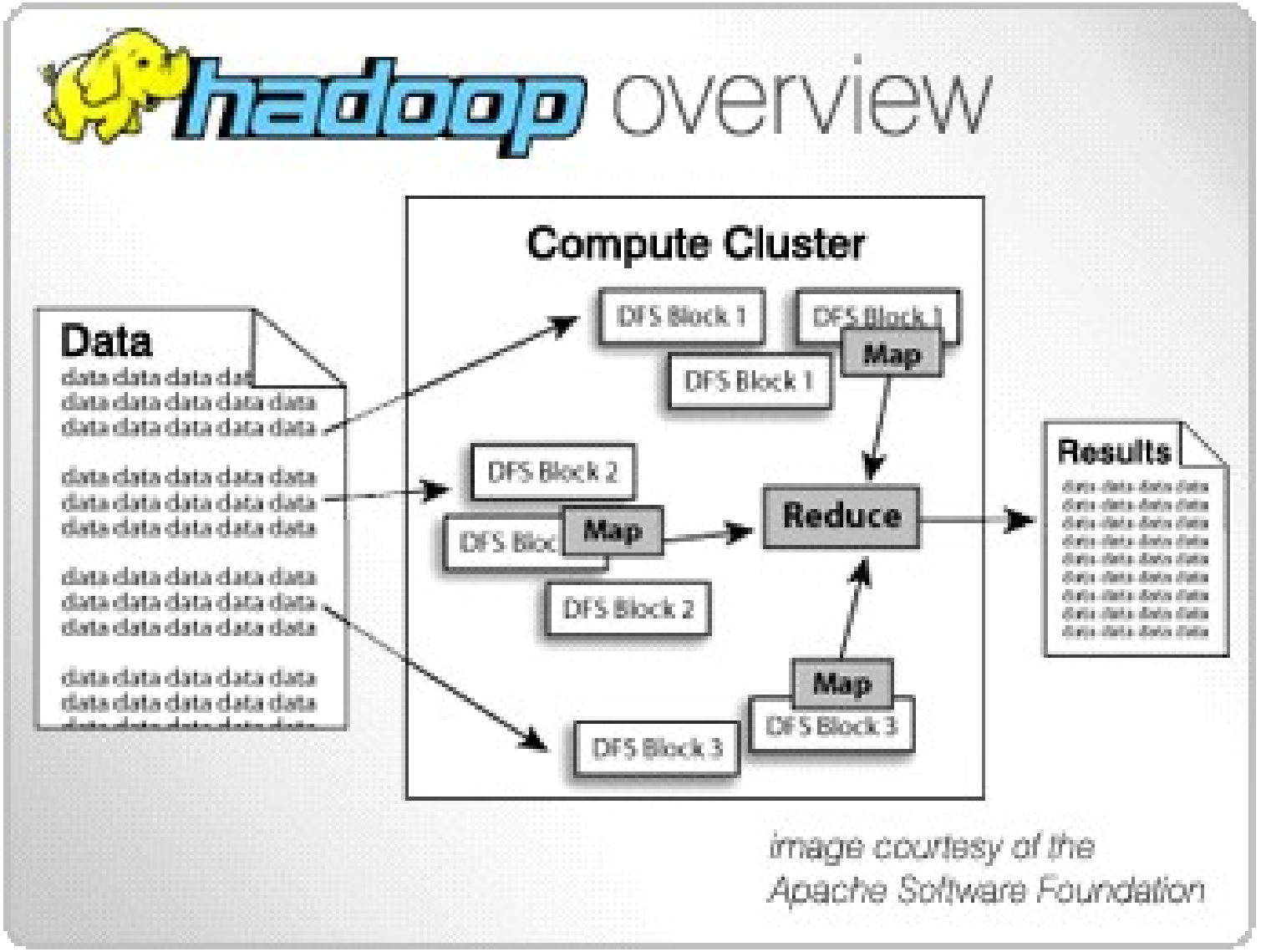
Large data: 17.5Gb every day – 10% Twitter

- input files are compressed splittable .lzo

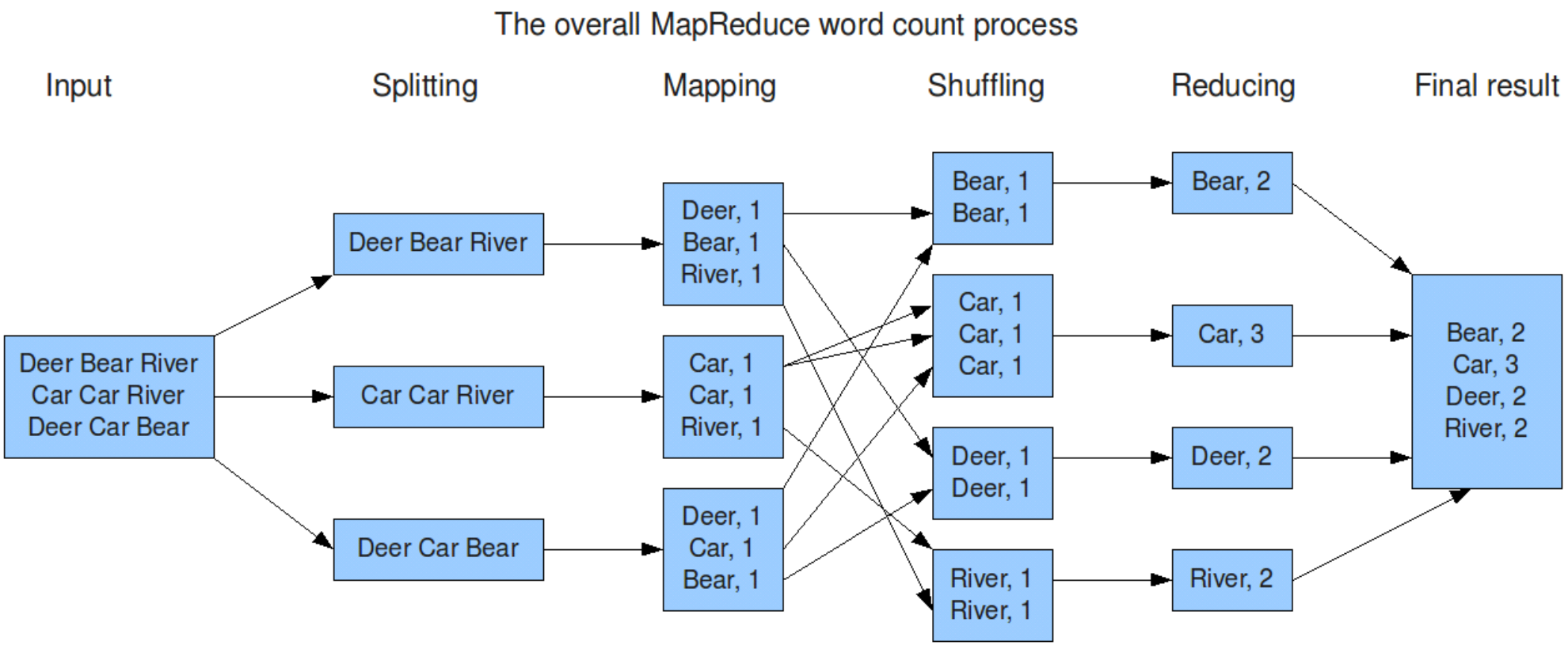
Many tasks can be done independently to each tweet

Run in parallel using Apache Hadoop Map-Reduce framework and distributed file-system

Architecture



Map Reduce Example



http://www.searchworkings.org/blog

Our Tool

Command line tool:

- single node
- distributed

2 types of usage:

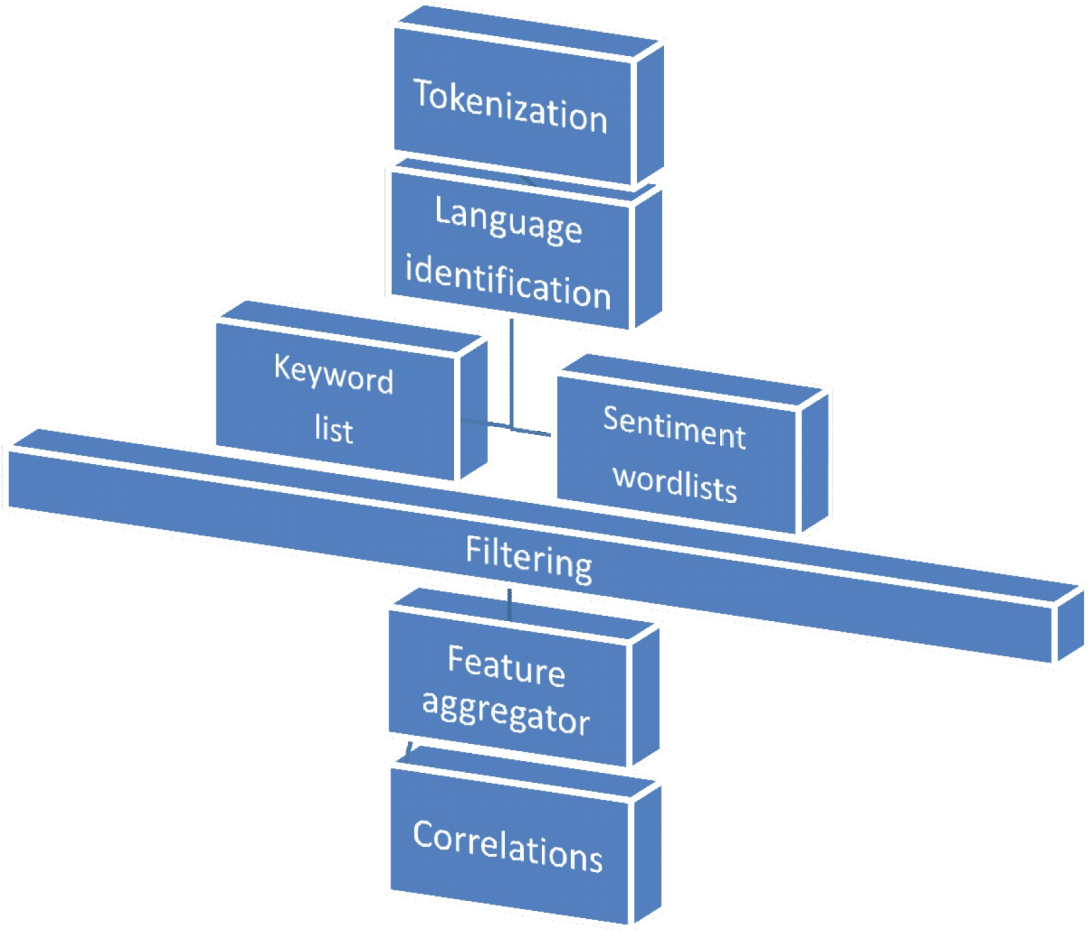
- online
- batch analysis

Scalable:

- can add new processing power in time

Use case

Mapper



Reducer

Regression models of trends in streaming data – Samangooei et. al. (2012)

Data format

- Twitter data comes in JSON format, so we also use JSON internally

- each step in the pipeline adds new fields to the record in a special "analysis" field

- supports USMF (Unified Social Media Format) developed by Tawlk

Data format

Input:

{…,
"text":"RT @MediaScotland greeeat!!! lvly speech by cameron on scott's indy :) #indyref",
"user":{"screen_name":"abx1","location":"sheffield,uk","utc_offset":0" …},
…}

Output:

{…,
"text":"RT @MediaScotland greeeat!!! lvly speech by cameron on scott's indy :) #indyref",
"user": {"screen_name:"abx1",[…]},
"analysis":{
"tokens":
["RT","@MediaScotland","greeeat","!!!","lvly","speech","by","cameron","on","scott's","indy",":)","#indyref"],
"ner": ["MediaScotland","cameron","scott's"],
"pos": ["~","@","^","","","A","N","P","^","P","L","N","E","#"],
"spam":"false",
"geo": {"city": "Sheffield", "country":"England", "long":"-1.46", "lat":"53.38", "population":"534500"},
"langid": {"language:" "en", "confidence": 0.51} }

Tokenizer

- Developed our own Twitter-specific tokenizer
- Works through a chainable set of regular expressions
- Can handle:
    - URLs
    - strange usage of punctuation
    - emoticons
    - hashtags, retweets, @ mentions
    - abbreviations, dates
- Currently only works for Latin scripted languages
- provides 2 outputs: protected and non-protected

Tokenizer

Example

Tweet: "@janecds RT badbristal np VYBZ KARTEL - TURN & WINE< WE DANCEN TO THIS LOL?http://blity.ax.lt/63HPL"

Tokens: [@janecds, RT, badbristal, np, VYBZ, KARTEL, -, TURN, &, WINE, <, WE, DANCEN, TO, THIS, LOL, ?, http://blity.ax.lt/63HPL]

Language detection

Detect language automatically (assume one language/tweet) and don't rely on user's self-reported profile language

We have reimplemented Lui and Baldwin's (2011) language detector - fast, standalone, pre-trained, 97 languages, different scripts

Test data: 2000 tweets in 5 languages from (Carter et al. 2012)

TextCat (5-way, raw)	TextCat (5-way, non-pr)	Lui & Baldwin (97-way,non-pr)
80%	89%	89.3%

Stemming

Using the Porter stemmer

Example

Tweet: "Tonight is the night!!Who is going to watch the second Semi-Final with us?? Got any crazy parties planned?"

Tokens: "Tonight is the night Who is going to watch the second Semi Final with us Got any crazy parties planned"

Filtering

Filter tweets based on values of attributes

Examples

- geo-tagged tweets

Have non-empty 'place' or 'geo' fields

- tweets with smileys

Have ':)' in their token list

- tweets that are pushed from Foursquare

Have 'foursquare' as their source

Geolocation

Map a tweet to it's sender geo information

At the moment: based on parsing the location field and timezone,

UK only

Example

“location”: “alton”,“utc_offset”: “0”

"geo": {
"city": "Alton",
"county": "South East England",
"lat": "51.14979934692383",
"population": "16584",
}
"country": "England",
"db_link": "http://dbpedia.org/resource/Alton,_Hampshire",
"long": "-0.9768999814987183",
"region": "SOU"

Analysis/Machine Learning

Word / Feature counts
Ex: For time series analysis

Pointwise Mutual Information (PMI) (exact and randomized versions)
Ex: Word co-occurrence analysis over time

Linear regression
Ex: For sentiment classification

Real time processing

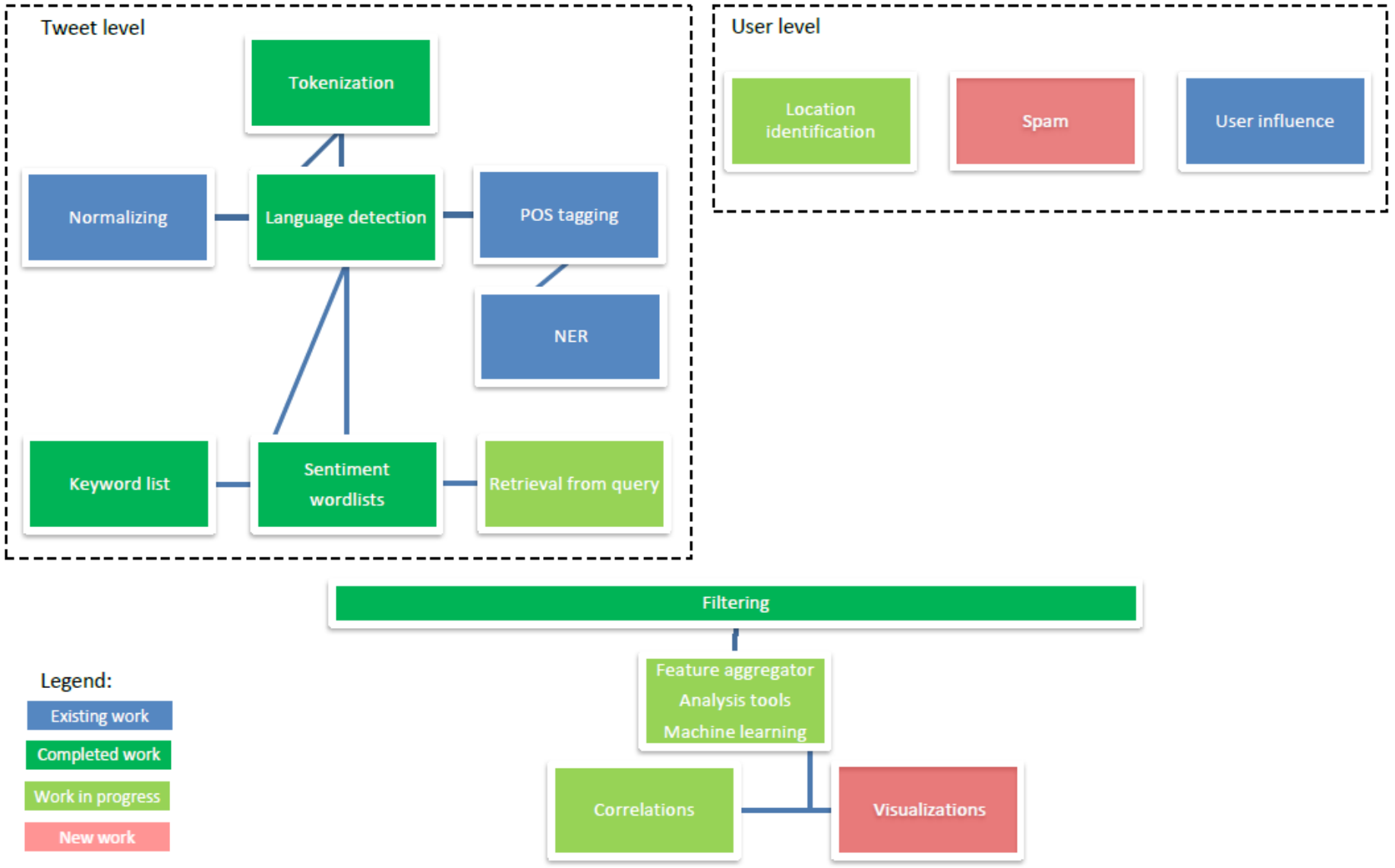
No. of tweets (in millions) processed (tokenized and language detected) in 1 hour:

Tw. Gardenhose (10% as of March 2012)	Single Core	Hadoop cluster
1.1	0.5	16

Pipeline can work in an online setting

* Hadoop cluster: 6 machines with 42 physical cores, max. 84 map tasks in parallel

Future plans



Future plans

Part-of-Speech tagging [Gimpel et al., 2011]

RT/~ @MediaScotland/@ greeeat/^!!!/,lvly/A speech/N by/P cameron/^ on/P scott's/L indy/N :)/E #indyref/#

Named entity recognition [Ritter et al., 2011]

RT @MediaScotland greeeat!!!lvly speech by cameron on scott's indy :) #indyref

Text Normalisation [Han & Baldwin, 2011]

RT @MediaScotland greeeat (great)!!!lvly (lovely) speech by cameron on scott's indy (independence) :) #indyref

User influence

Using the Klout API, gives a score from 0-100 to each OSN user.

KLOUT

More information

"Trendminer: An Architecture for Real Time Analysis of Social Media Text"

[Preotiuc-Pietro D., Samangooei S., Cohn T., Gibbins N., Niranjan M.]

Real-Time Analysis and Mining of Social Streams (RAMSS) ICWSM 2012

Download and contribute (BSD license):

http://github.com/sinjax/trendminer

http://www.trendminer-project.eu

Deliverable 3.1.1 – Regression models of trends in streaming data

Thank you!