Thematic tweets categorization

Gauthier Pironi

Vendredi 27 février 2015

Supervision : Sylvie Ratté

Menu of the day

Context

Goals

Requirements

State of the art

Methodology

Implementation

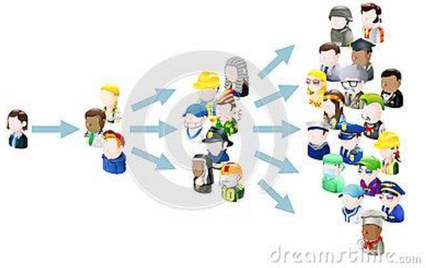
Results

Discussion

Conclusion

Context

- public relation company
- increase impact of press release
- concept of “category”



- the projet is a proof of concept

Goals

- thematically categorize the tweets
- 4 categories :
    -  Culture
    -  Beauty/Fashion
    -  Food
    -  Other
- Precision more important than recall
    - huge amount of data
    - having less categorized tweets is not that bad

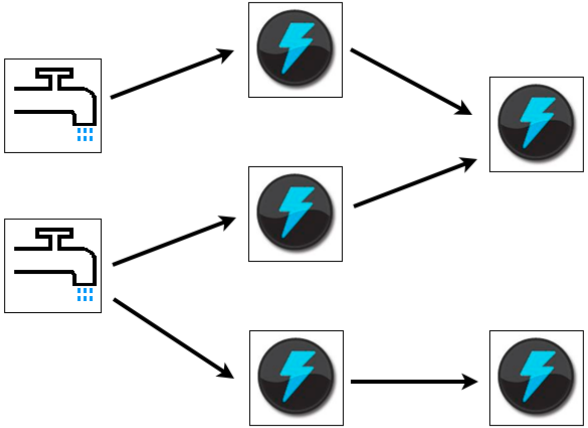
Requirements and constraints

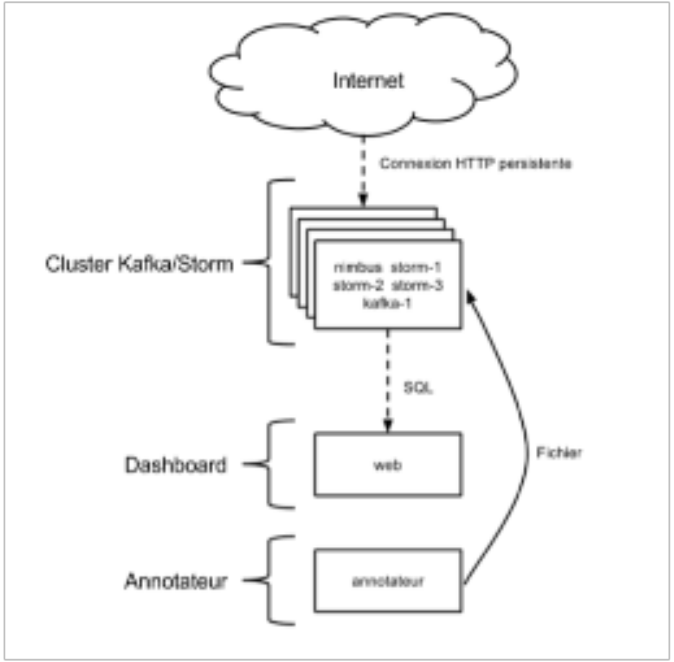
- Storm environment
  - →“big data”
- Java
- getting the data
- performance
  - ressources used by the categorization
  - results of the categorization
- real time!

Storm

- Distributed computation framework
- Nathan Marz
- first release in 2011
- Topology :
  - Spouts
  - Bolts

Storm - Architecture





State of the art

- classical problematic in NLP
- many fields : medical, understanding behavior, discovering trends, sentiment analysis, etc.
- 5 mains steps :
  - getting the data
  - normalization/cleaning the data
  - extract features
  - statistic tool - classifier
  - model validation

State of the art - normalization

- Tokenization
- PoS-Tagging
- Stop words (the, a, this, her, his, etc.)
- Named-entity recognition (NER)
- Stemming

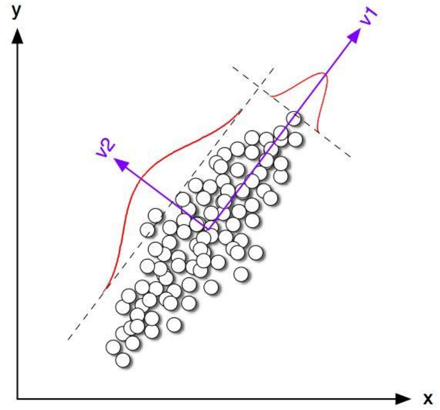
Racine	Exemples de mots possédant la racine
anim	animal, animal's, animality, animals, animate, animatedly, animates, animating, animation, animation's, animations, animator, animator's, animators, anime, anime's, animism
inform	informal, informality, informally, informals, informant, informant's, informants, information, information's, informational, informations, informative, informatively, informer, informer's, informers, informs
liber	liberal, liberal's, liberalism, liberalism's, liberality, liberalization, liberalizations, liberalize, liberalizes, liberally, liberals, liberate, liberates, liberation, liberation's, liberator, liberators

Tableau 4.1 Exemples de 3 racines et de mots possédant ces racines.

State of the art - extracting features

- N-gram
- tf-idf

$$tf-idf_{t,d} = tf_{t,d} \times idf_t = tf_{t,d} \times log(\frac{N}{df_t})$$

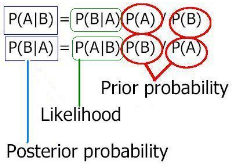


- meta features (author, keywords, dictionnaries, etc.)
- dimension reduction (PCA, chi-squared test, etc.)

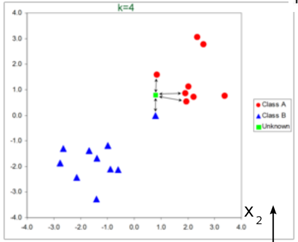
State of the art - classifier

Supervised

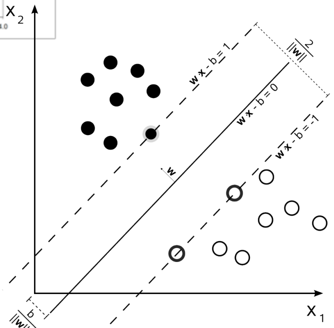
-naive bayes



-kNN



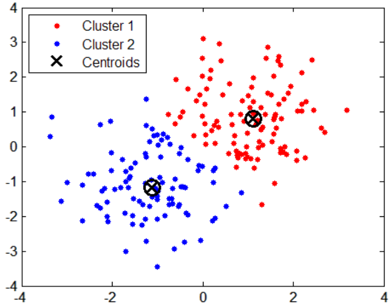
-SVM



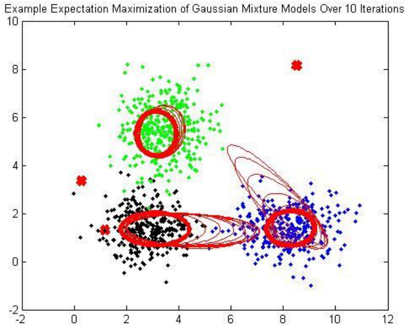
vs.

Unsupervised

kmeans-



EM-



State of the art - Results

Article	Comment	Results
Fürnkranz (1998)	- n-gram - ~10 000 features.	P : 80% R : 82%
Cano *et al.* (2013)	-model for violence detection in social media -binary output y/n	P : 82% R : 89%.
Horn et Center (2009)	-3 categories : user, company and news	P(news) : *83%* P(user) : *84%* P(company) : *78%*

P = Precision / R = Recall

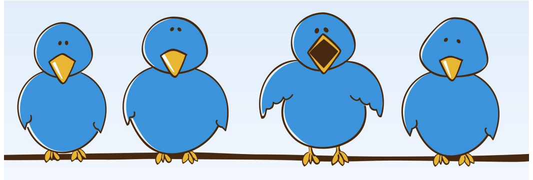
Methodology

Chosen approach in 5 steps:

- clean / normalize the data
- n-gram + additionnal features
- supervised classifier
  - golden corpus
  - annotator
- Model evaluation
- Optimization

Methodology

Let's take a first look at the data : tweets...



For The Love Of Christ, Don't Book A Hotel Without First Consulting @AHotelLife here's why http://t.co/vnsulZcSfE

@habituallychic Forget Frenchmen, I'd rather make out with a macaron #reallythough

RT @Thezog : I think she won for starting and not finishing the most sentences #Bissett #Goldenglobes

The train to Montauk has so many Irish students on it I feel like I'm on Iarnrod Eireann. Minus the tea trolley sadly.

Not sure your brain can handle this one : Baby pandas on a fucking SLIDE http://t.co/oKeQzvml

Tableau 5.1 Exemples de tweets provenant du Canada anglophone.

Methodology - Gold corpus

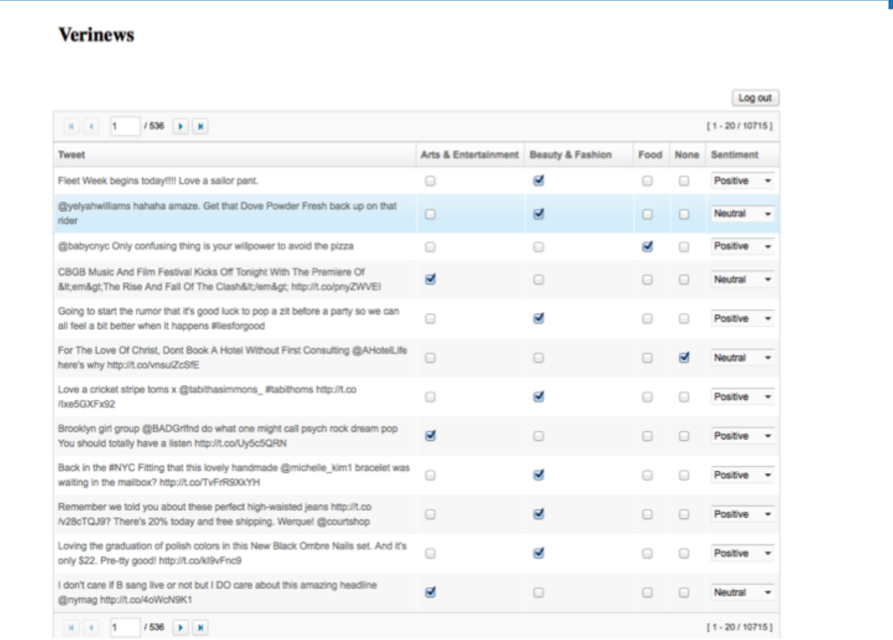
- Supervised classifier → annotated tweets
- custom categories → create our own corpus

- obtaining the data:
    - 36 Twitter accounts over a month→ ~120 000 tweets
    - 10% random tweets from each user → 1100 tweets

Thanks to the UTPL team in Ecuador (and Sylvie)!

Methodology - Gold corpus

- tweets annotation
- web annotator for the client



Catégorie	Nombre de tweets	Proportion
Culture	90	16,57%
Mode/Beauté	130	23,94%
Cuisine	59	10,87%
Autre	264	48,62%

Tableau 5.2 Repartition des donnes selon les différentes catégories.

- 543 annotated tweets in ~3 months

Methodology - normalization

- Java Regex
  - powerful
  - friendly and fun to use
  - easy to optimize
  - very good performance! (CPU + memory)
- String methods
  - toLowerCase()

Opération effectuée	Texte avant ⇒ après l'opération
Remplacer les utilisateurs par *username*.	@naagofficial ⇒ username
Retirer les caractères # des hashtags.	#sharkattack ⇒ sharkattack
Remplacer les URL	http ://t.co/k8NpvAO3yY ⇒ url
Retirer les mentions de retweet : RT	RT i love it ⇒ i love it
Retirer les chiffres	sold 30 million ⇒ sold million
Mettre le texte en minuscule	NY WEEK diary ⇒ ny week diary

Tableau 5.3 Les différentes opérations de nettoyage du texte.

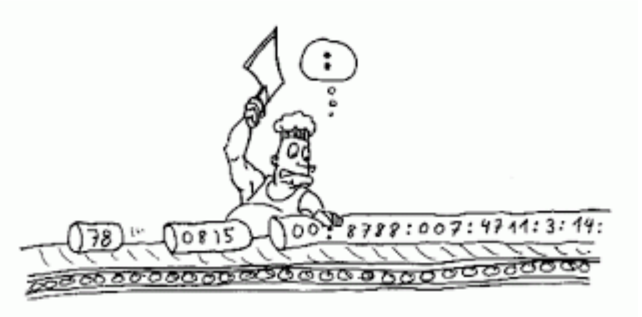
```
// remove @username
tmpTweet = tmpTweet.replaceAll("@\\S*", " username ");

// replace #hashtag with hashtag
tmpTweet = tmpTweet.replaceAll("#", "");
```

```
// replace http://myurl.com with url
tmpTweet = tmpTweet.replaceAll("https?://[\\S]*", " url ");
// remove ponctuation .,;:!?"*
tmpTweet = tmpTweet.replaceAll("[.:,;()\"*]", " ");
```

```
tmpTweet = tmpTweet.toLowerCase();
```

Methodology - features

- tokenizer on white space
- removing stop words
- stemming



http://alifewhatever.blogspot.ca/2011/11/java-string-tokenizer-example.html

a	ourselves
about	out
above	over
after	own
again	same
against	shan't

http://www.ranks.nl/stopwords

→ unigrams with high and low threshold

Methodology - meta features

- In addition to “classical” features
- based on dictionnaries (list of thematic words)

Catégorie	Nombre de mots	Exemples de mots du dictionnaire
Culture	94	sculptor, sculpture, sewing,shows, singer
Beauté/Mode	1461	hair, hairstyle, makeup, beauty, skin
Cuisine	550	pumpkin, punch, quiche, quinoa, radish

Tableau 5.4 Récapitulatifs des 3 dictionnaires

- created dictionnaries manually (mainly from wikipedia)
- 4 meta features

Methodology - meta features

Meta feature 1 : tweets content

Tweet
By the way, would you let our server know that when we asked for chai, he brought us a cup of hot water ? AMAZING. http://t.co/EM10Sc40

Valeur des méta-attributs
Culture : 0
Beauté/Mode : 1
Cuisine : 4

Tableau 5.5 Exemple de tweet avec la valeur des 3 méta-attributs qui portent sur le contenu de celui-ci.

Methodology - meta features

Meta feature 2 : URL link

Lien t.co généré par Twitter http://t.co/4S5DUaZ2
Vrai lien après redirection http://intothegloss.com/2013/01/ren-glycol-lactic-radiance-renewal-mask/
Liste de mots extraits du lien intothegloss, 2013, 01, ren, glycol, lactic, radiance, renewal, mask
Valeur des méta-attributs Culture : 0 Beauté/Mode : 4 Cuisine : 0

Tableau 5.6 Exemple de lien hypertexte avec les valeurs des 3 méta-attributs qui portent sur les mots composant le lien.

Methodology - meta features

Meta feature 3 : meta content in the page (HTML’s <meta> tags)

Lien t.co généré par Twitter http://t.co/0MKyS3awiW
Vrai lien après redirection http://byrnenotice.com/michelle-siwys-newest-wildfox-denim-collection-solidifies-her-master-of-denim-status/
Extrait des balises <meta> de la page pointée par le lien The Byrne Notice is a fashion, lifestyle and culture site featuring travel, beauty, food, nightlife, books, art, interiors and more, all through the lens of fashion of downtown New York City. It includes interviews, guides to the coolest and newest places and faces, party galleries and more. Fashion, Beauty, Culture, Lifestyle, Hipster, New York, Downtown, Williamsburg, Style, Food, Best Restaurant, Best Bar, Nightlife, Interiors, Home Decor, Home Design, Accessories
Valeur des méta-attributs Culture : 4 Beauté/Mode : 6 Cuisine : 4

Tableau 5.7 Exemple de lien hypertexte avec les valeurs des 3 méta-attributs qui portent sur les meta-données de la page pointée par le lien.

Methodology - meta features

Meta feature 4 : text content in the page (HTML’s <p> tags)

Lien t.co généré par Twitter
http://t.co/pnyZWVEI

Vrai lien après redirection
http://byrnenotice.com/cbgb-music-and-film-festival-kicks-off-tonight-with-the-premiere-of-the-rise-and-fall-of-the-clash

Extrait des balises paragraphes <p> de la page pointée par le lien
They’re doin’ it right, too. There are over 300 bands scheduled, and not Ramones cover acts either ; skimming the list, we saw Clap Your Hands Say Yeah, Guided By Voices, Pains of Being Pure At Heart, Cloud Nothings, War on Drugs, The Virgins, Lissy Trullie, Dale Earnheardt Jr Jr, and one of our new ar...And if you’re really into CBGB and movies, you’ll be excited to hear they started shooting CBGB : The Movie last month (odd fact : “rabid punk fan” Rupert Grint, of Ron Weasley fame, is playing the guitarist for favorite CBGB band Dead Boys. It’s okay, we don’t know, either)

Valeur des méta-attributs
Culture : 7
Beauté/Mode : 3
Cuisine : 5

Tableau 5.8 Exemple de lien hypertexte avec les valeurs des 3 méta-attributs qui portent sur le texte pointé par le lien.

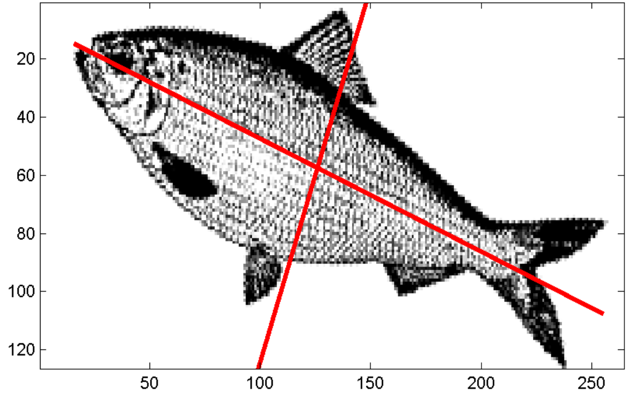
Methodology - dimension reduction

PCA

```
 possibly correlated variables
               ↓
linearly uncorrelated variables
```

$$\mathbf{V}^{-1}\mathbf{C}\mathbf{V} = \mathbf{D}$$

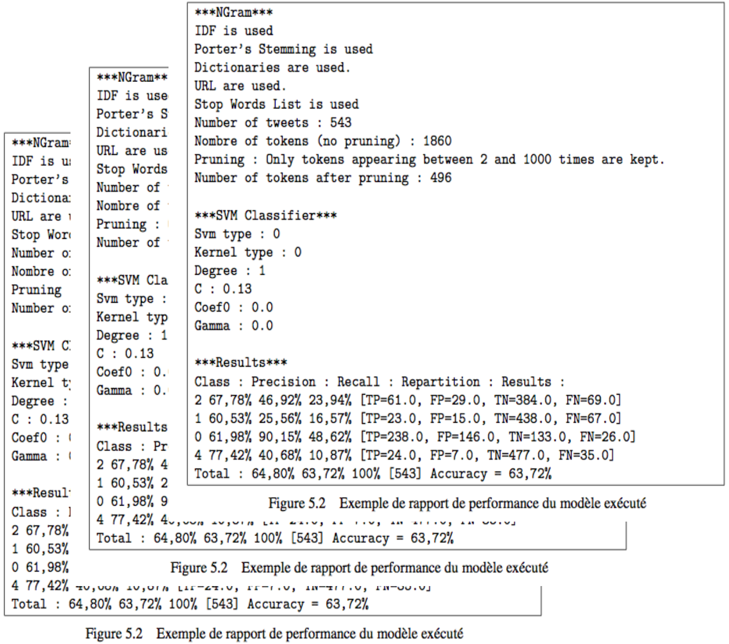
- reduce the number of features
- could increase performance:
  - classifier results
  - ressources usages



Methodology - classifier & validation

- Classifiers:
  - SVM
  - Naive Bayes
  - kNN
  - zeroR (majority class) for baseline
- Validation
  - cross-validation
  - 10 folds

Classifier optimization



```
***NGram***
IDF is used
Porter's Stemming is used
Dictionaries are used.
URL are used.
Stop Words List is used
Number of tweets : 543
Nombre of tokens (no pruning) : 1860
Pruning : Only tokens appearing between 2 and 1000 times are kept.
Number of tokens after pruning : 496

***SVM Classifier***
Svm type : 0
Kernel type : 0
Degree : 1
C : 0.13
Coef0 : 0.0
Gamma : 0.0

***Results***
Class : Precision : Recall : Repartition : Results :
2 67,78% 46,92% 23,94% [TP=61.0, FP=29.0, TN=384.0, FN=69.0]
1 60,53% 25,56% 16,57% [TP=23.0, FP=15.0, TN=438.0, FN=67.0]
0 61,98% 90,15% 48,62% [TP=238.0, FP=146.0, TN=133.0, FN=26.0]
4 77,42% 40,68% 10,87% [TP=24.0, FP=7.0, TN=477.0, FN=35.0]
Total : 64,80% 63,72% 100% [543] Accuracy = 63,72%
```

Figure 5.2 Exemple de rapport de performance du modèle exécuté

→ Finding the best combination of tools to use in the model

→ Finding the best meta parameters for each classifier

Classifier optimization

What combination of tools is the best one?

- 6 tools :
  - threshold for n-gram (min-max)
  - stop words
  - stemming
  - tf-idf
  - meta features
  - dimension reduction

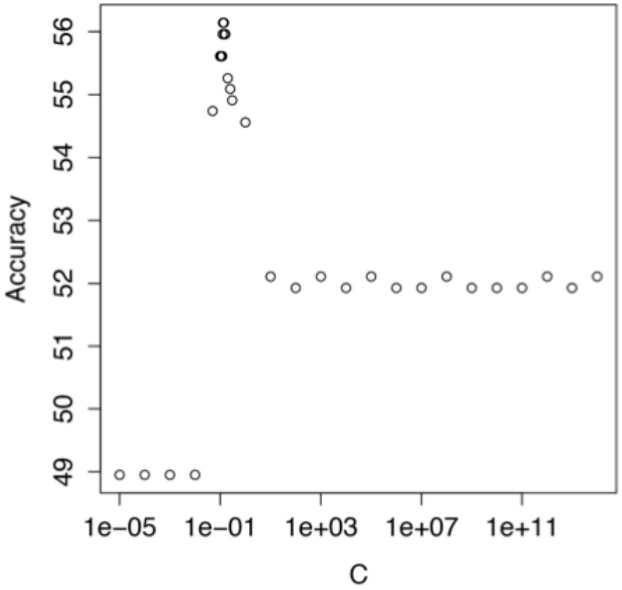
Classifier optimization

What are the best meta parameters for each classifier?

(→What is the best classifier?)

- SVM
- kNN
- zeroR
- naive Bayes

Classifier optimization - Linear SVM



- Only 1 parameter "C" to optimize here
- Optimization criteria : accuracy
- Grid Optimization

→ First :

From C = 10^−10 to C = 10^5 with a multiplying factor of 100

→ Second

C = 10^−10 to 10^−9 with a multiplying factor of 10

→ return to first step

Classifier optimization - RBF kernel SVM

- 2 parameters : C,γ (gamma)

→“Heatmap” graph

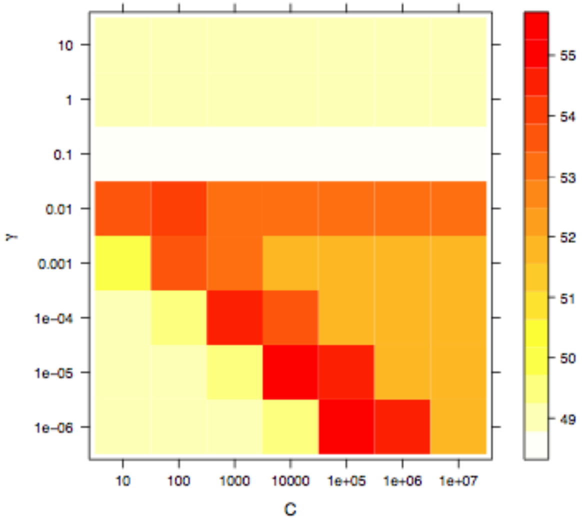


Figure 5.4 Graphe d’optimisation d’un classifieur avec un noyau RBF et ses deux paramètres C et γ. L’échelle de couleur représente la proportion de tweets bien classifiés : plus la case est rouge, meilleure est la combinaison des paramètres $\{C, \gamma\}$ correspondante.

Classifier optimization - sigmoid kernel SVM

- 3 parameters : *C*,γ (gamma), coef0

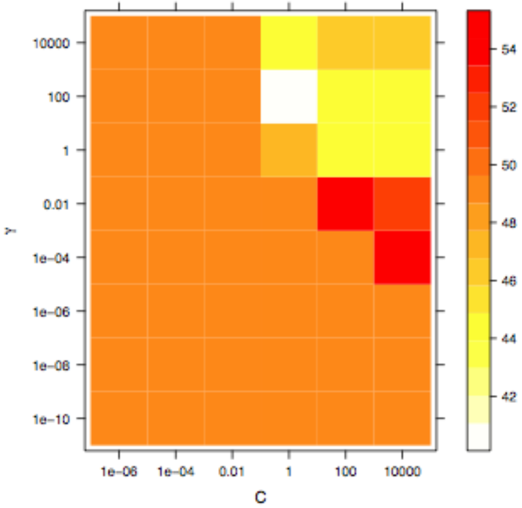


Figure 5.5 Graphe d'optimisation d'un classifieur avec un noyau sigmoïde et deux de ses trois paramètres C et γ. Afin de représenter graphiquement l'optimisation, le paramètre $coef0$ a été ici fixé à une valeur de 0.

Implementation

- All the categorization is done in Java
    - in a *bolt* (the Storm logical processing unit)
- Regex in Java for cleaning the tweets
- Snowball stemmer
- URL reading : jsoup
    - great library! But reading URL is slow….
- Java Machine Learning Library (JavaML)

Results

Mots stemmés	Frequence
usernam	462
url	275
!	159
?	128
love	37
dai	17
good	15
&	13
amaz	12
feel	12
time	12
todai	12
beauti	10
happi	10
babi	8

Tableau 7.1 Mots stemmés les plus fréquents dans le corpus.

- Most frequent stemmed tokens:
  - username, url, etc.
  - love, good, feel, time, beauty, baby, etc.

Results - Best combination of tools

- n-gram threshold
  - min = 2
  - max = 500
- stop words
- stemmer
  - Stemmer snowball
- tf-idf
- meta features (all of them!)

→ Using all the tools **except PCA** is the best combination

Results

The best one! →

	Culture		Beauté/Mode		Cuisine		Autre		
Modèle évalué	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**Acc.**
1G + MA + zeroR	00.00	00.00	00.00	00.00	00.00	00.00	48.62	**100**	48.62
1G + MA + kNN[a]	14.29	01.11	49.04	39.23	80.00	33.90	55.53	85.61	54.88
1G + MA + Bayésien naïf	56.25	10.00	59.74	35.38	62.50	08.47	55.43	92.80	56.17
1G + MA + SVM-S	47.06	**35.56**	49.62	**50.77**	64.10	**42.37**	61.39	70.45	56.91
1G + MA + SVM-RBF	57.45	30.00	61.76	48.46	67.57	**42.37**	**62.75**	84.85	62.43
1G + MA + SVM-L	**57.89**	24.44	**66.67**	46.15	**77.42**	40.68	61.46	89.39	**62.98**
1G + MA + SVM-L + ACP	41.89	34.44	[illegible]	48.46	50.00	38.98	62.26	[illegible]	57.09
1G + SVM-L	39.53	18.89	58.82	30.77	52.94	15.25	54.94	86.36	54.14
MA + SVM-L	55.00	24.44	65.22	34.62	66.67	37.29	58.35	88.64	59.48

P : précision
R : rappel
Acc. : proportion globale de tweets bien classés (*Accuracy*)
1G : unigramme
MA : méta-attributs (dictionnaires)
SVM-S : SVM à noyau sigmoïde
SVM-RBF : SVM à noyau à base radiale
SVM-L : SVM à noyau linéaire

a. On utilise $k = 3$
b. On conserve $n = 200$ composantes principales

Results

								
P	R	P	R	P	R	P	R	Acc.
57.89	24.44	66.67	46.15	77.42	40.68	61.46	89.39	62.98

P = Precision
R = Recall
Acc. = Accuracy

Discussion

- increase the size of the golden corpus
  - requires efforts from the client!
- test new classifiers that have already proven good performance (C4.5)
- use subcategories for a finer categorization:
  - culture → movie, music, literature, theatre, etc.
  - beauty/fashion → haute couture, make up, shoes, etc.
- use dynamic dictionnaries

Discussion

- take into account the author of the tweet:
  - need to categorize the author also?
    - previous tweets
    - author’s profile (picture, description)
  - popularity
- take into account the images
  - images in the tweets or in URL pointed by the tweets
  - the user’s profile picture
  - use pattern recognition techniques

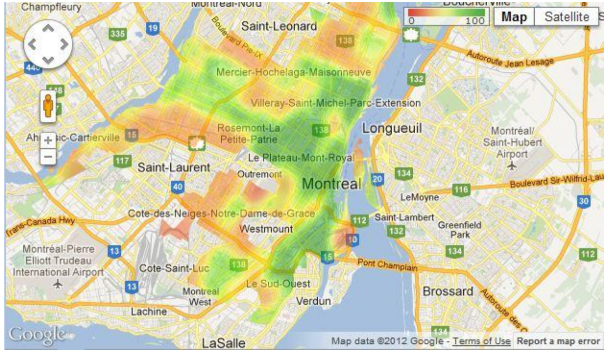
Conclusion

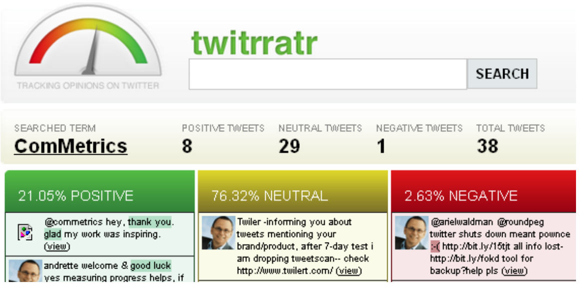
- satisfying categorization performance
- a successful first step in the project
- the categorization method and algorithm can be used for any domain
    - individual components (classifier, tweet cleaner,etc.) and parameters can be easily changed
- have tweets classified in multiples categories

Conclusion - Why categorize?

With the categorized tweets :

- keywords suggestions
- sentiment analysis
- trends
- filtering
- heatmaps





Thank you for listening!

Questions?