

OPITER : Fouille de données d'opinion pour les territoires

Eric Kergosien (TETIS- LIRMM)

Pierre Maurel (TETIS)

Mathieu Roche (TETIS– LIRMM)

Maguelonne Teisseire (TETIS– LIRMM)



Le projet Senterritoire



Géographes - Informaticiens



Laboratoire
d'Informatique
de Robotique
et de Microélectronique
de Montpellier

LIRMM



Informaticiens



→ proposer un environnement décisionnel basé sur une analyse automatique des textes liés à l'aménagement du territoire

Démarche générale

Documents d'actualités



Sommaire

1. La notion d'opinion
2. Chaîne de traitements standard de la communauté
3. L'approche OPITER
 - a. Démarche
 - b. Un petit exemple
 - c. Experimentations
4. Conclusion et travaux en cours

Qu'entend-on par Opinion?

- Opinion : négative, positive, neutre $\neq$ Sentiments : joie, tristesse...



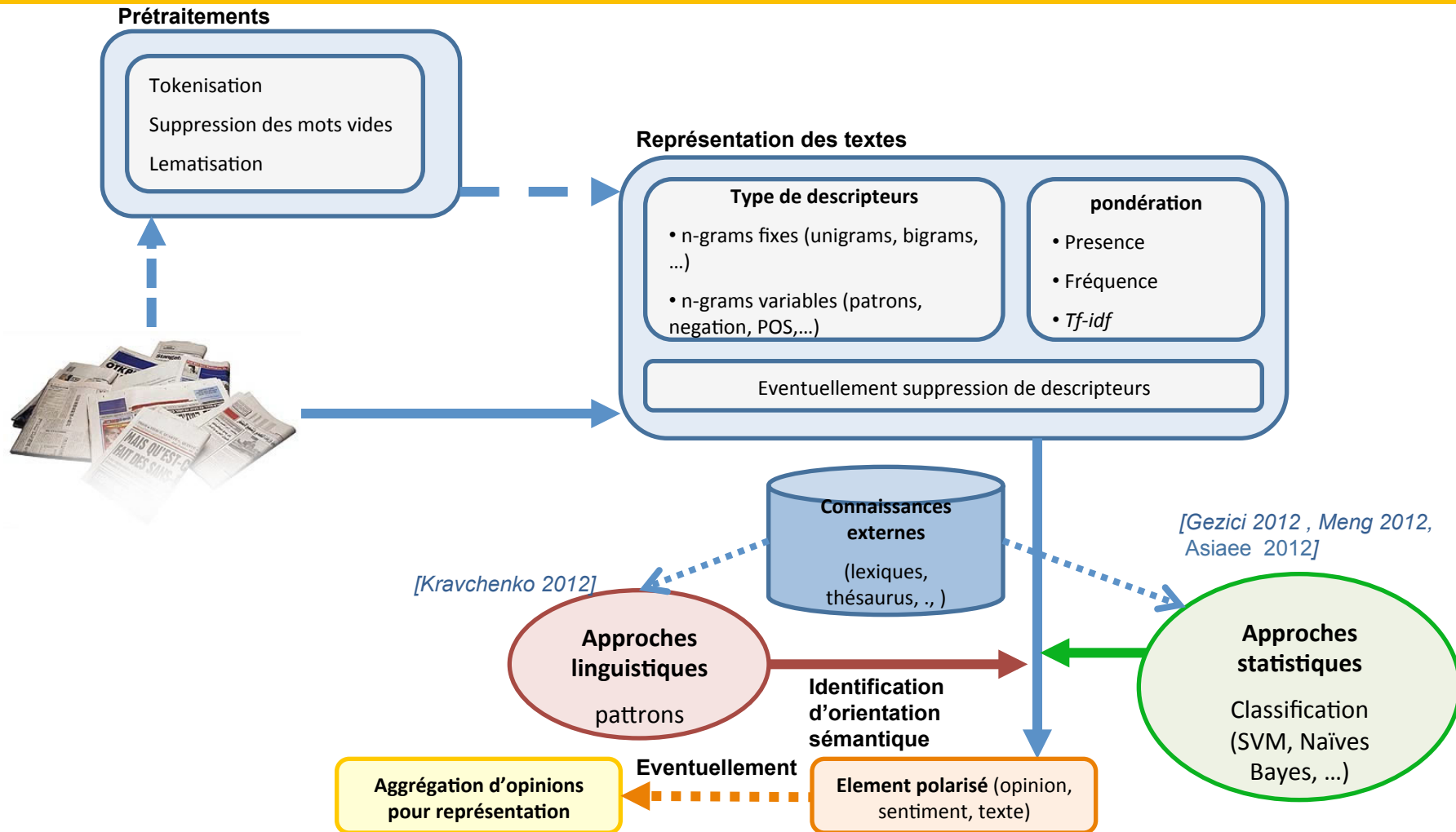
PROBLEMATIQUE



Quelles sont les spécificités des opinions relatives aux informations géographiques et comment les extraire ?

“J’aime beaucoup le centre ville de Montpellier et je déteste la plage de Carnon”

Chaine de traitements en fouille d'opinions



Pas de travaux pour la détection d'opinions dans l'aménagement du territoire
Proposition : tester les méthodes classiques pour évaluer nos ressources

Ressources

Data

SENT_100



99 Articles du *Midi libre* depuis 2006, organisés en deux classes par les experts géographes (positifs et négatifs)

SENT_150

150 articles du *Midi libre* organisés en deux classes par notre équipe (positifs et négatifs)

3 corpus DEFT

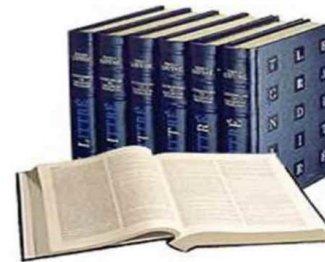
- 300 extraits d'un débat parlementaire de l'Assemblée nationale;
- 1000 Critiques de jeux vidéos;
- 1500 critiques de films.



Lexicons

General Inquirer

Version française (Bestgen 2011), 1246 mots positifs et 1527 mots négatifs



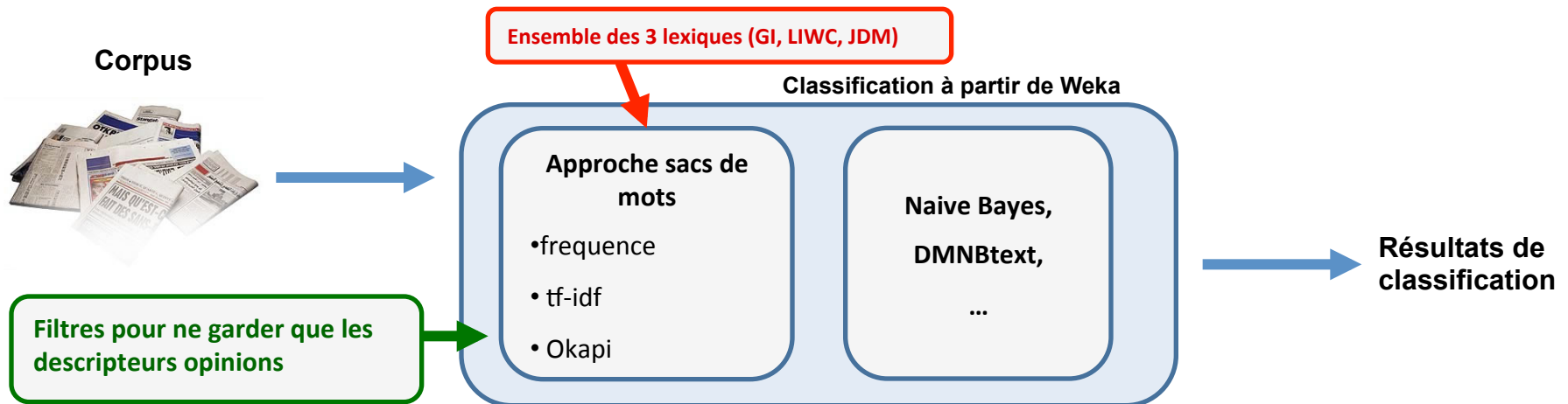
Jeux de mots (JDM)

9653 mots positifs et 6700 mots négatifs en français

LIWC

Version française (Piolat and al. 2011), sélection des termes positifs et négatifs

Premières expérimentations pour tester les lexiques d'opinion



	SENT_10 0	SENT_15 0	DEFT_300	JV_1047	AVAL_1459
DMNBtext	51.51%	56%	86.3%	93.65%	90.13%
Pondération des mots d'opinions	49.49%	54%	86.66%	93.95%	88.21%

- Approches classiques pas adaptées pour un corpus de petite/moyenne taille relatif à l'aménagement du territoire
- Pas d'amélioration des scores en donnant plus de poids aux lexiques d'opinion

Démarche générale

3. Identification des descripteurs représentatifs de polarité du domaine

2. Prise en compte du contexte des opinions pivots

1. Identification du lexique d'opinions pivots



Corpus Senterritoire

Vocabulaire d'opinions

Vocabulaire d'opinions

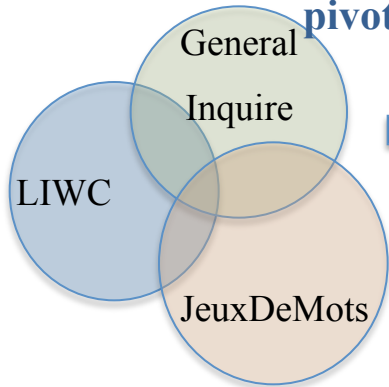


Vocabulaire d'opinion

pivots généraliste VPOG

contextualisé VOC

spécialisé VOS



Intersection de

lexiques d'opinion

Exemple

1 : positive The XXX basin is made up of the magnificent and natural XXX lagoon, its watershed and marine front.

2 : positive Increasing urbanisation means greater competition with vine-growing and natural space linked with the beautiful lagoon.

3: negative However, the XXX fears that fishermen of XXX will move elsewhere to find work. The heavily criticized General Fisheries Council for the XXX is working with the XXX in order to find a solution.

4: positive The XXX is a natural marine environment, unrivalled in XXX, so leisure activities are strictly controlled to protect this eco-sensitive zone.

....

100 : negative From 1880 to 1884, seven concessions were granted in the XXX canals. The consequences of these installations in heavily polluted waters did not take long to manifest, with gastrointestinal intoxications and even cases of typhoid fever.

Color legend:

- Negative opinion from GVPO: Expression
- Negative opinion from CVO: Expression
- Negative opinion from SVO: Expression
- Positive opinion from GVPO: Expression
- Positive opinion from CVO: Expression
- Positive opinion from SVO: Expression

Extraction des mots polarisés relatifs au domaine

Attribution d'un score d'opinion par document pour évaluation

Nettoyage (on ne garde que les noms communs, adjectifs, verbes et adverbes)

$$\text{scoreMot}(\text{MotPivot}) = S_{i,l} \times \text{polarite}(\text{MotPivot}, l)$$

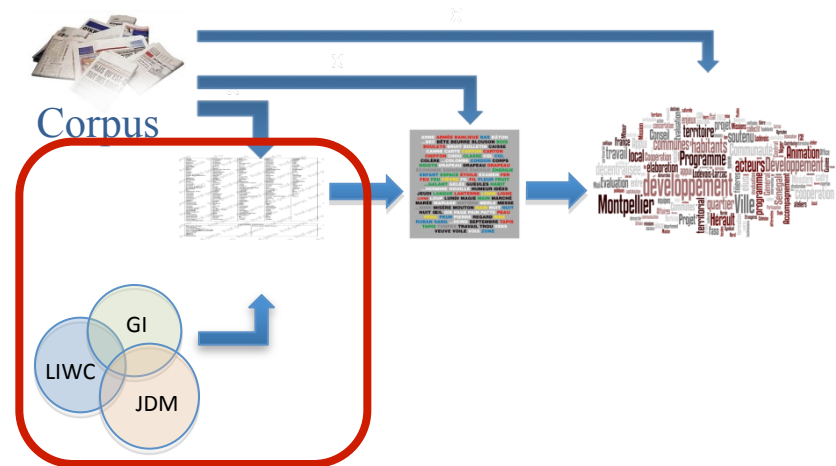
$$\text{scorePhrase}(P) = \frac{\sum \text{Cat} \times \text{scoreMot}(M)}{\sum \text{Cat} \times M}$$

$$\text{scoreDocument}(D) = \frac{\sum \text{scorePhrase}(\text{Phrase})}{\sum \text{Phrase}}$$

Lexique	Score de classification correcte
<i>GeneralInquirer</i>	57,5%
<i>LIWC</i>	54,5%
<i>JeuxDeMots</i>	51,5%

Extraction des mots polarisés relatifs au domaine

- $\text{GeneralInquirer} \cap \text{LIWC} \cap \text{JeuxDeMots} = S_1$;
- $\text{GeneralInquirer} \cap \text{LIWC} = S_2$;
- $\text{GeneralInquirer} \cap \text{JeuxDeMots} = S_3$;
- $\text{LIWC} \cap \text{JeuxDeMots} = S_4$;
- $\text{GeneralInquirer} = S_5$: Mots du GeneralInquirer distincts des autres lexiques ;
- $\text{LIWC} = S_6$: Mots du LIWC distincts des autres lexiques ;
- $\text{JeuxDeMots} = S_7$: Mots du JeuxDeMots distincts des autres lexiques.

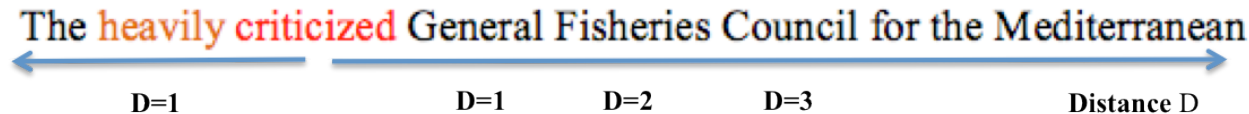
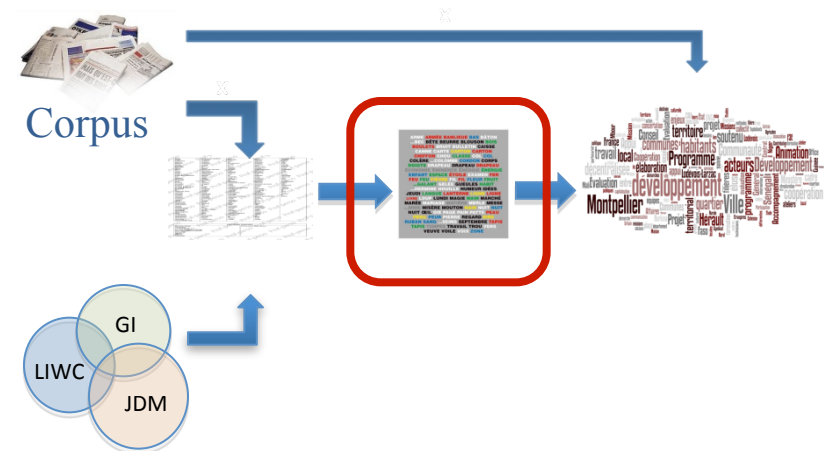


Evaluation du vocabulaire d'opinions pivots généraliste pour la classification de textes

Tests	GI-LIWC-JDM (S_1)	GI-LIWC (S_2)	GI-JDM (S_3)	LIWC-JDM (S_4)	GI (S_5)	LIWC (S_6)	JDM (S_7)	Scores
1	1	1	1	1	1	1	1	52,5%
2	1	1	1	1	0	0	0	59,6%
3	3	0	2	0	1	0	1	54,5%
4	3	0	0	2	0	1	1	54,5%
5	3	2	0	0	1	1	0	63,6%
6	3	2	0	0	1	0	0	64,6%

Extraction des mots polarisés relatifs au domaine

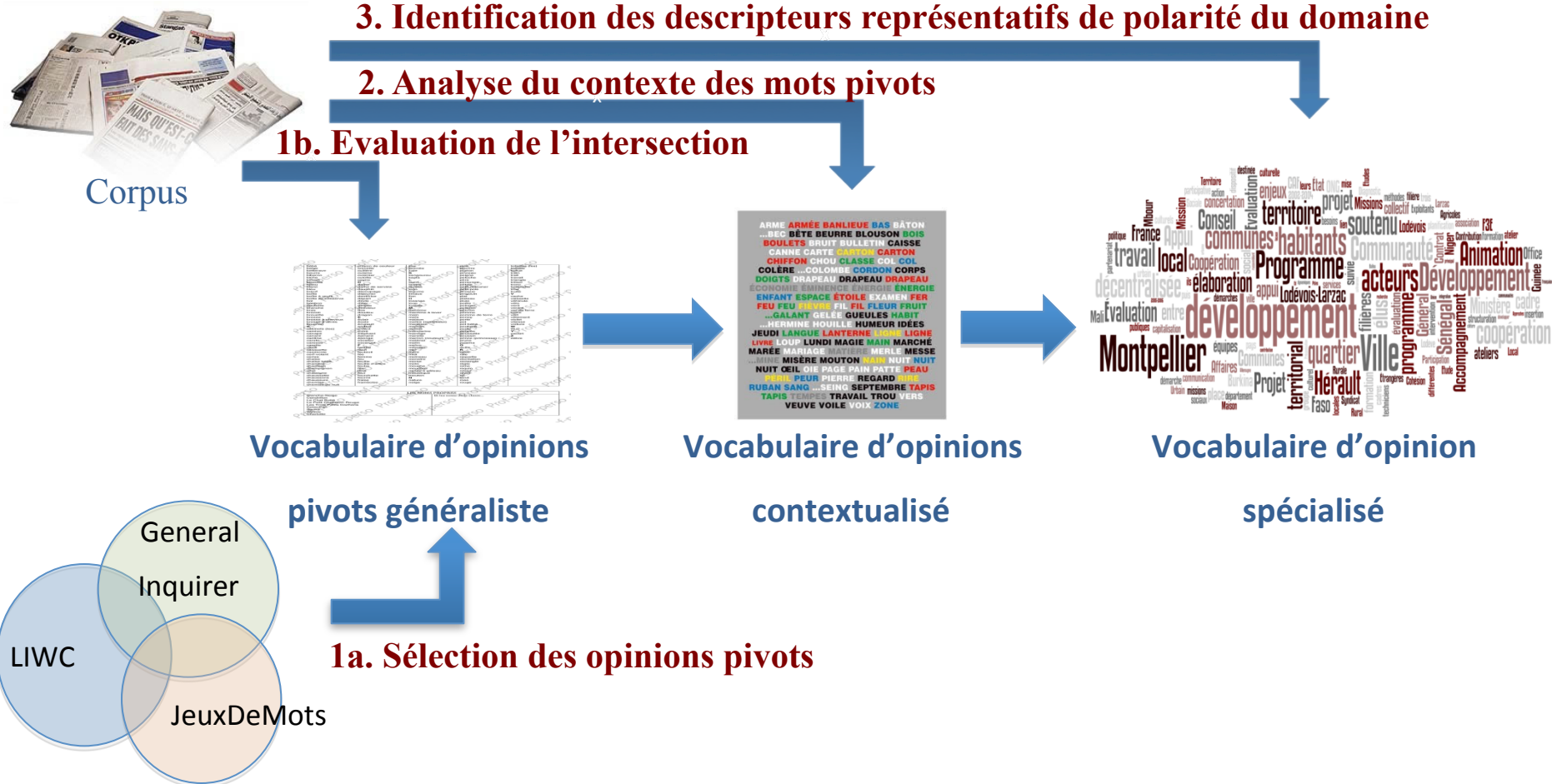
$$\text{ScoreMotVoisin}(\text{Mot}) = \frac{\sum \frac{\text{ScoreMot}(\text{MotPivot})}{d}}{\sum \text{MotsPivots}}$$



Evaluation du vocabulaire d'opinions contextualisé regroupant opinions pivots et mots voisins

Windows	GI-LIWC-JDM	GI-LIWC	GI-JDM	LIWC-JDM	GI	LIWC	JDM	Scores
0	3	2	0	0	1	0	0	64,6%
1	3	2	0	0	1	0	0	64,6%
4	3	2	0	0	1	0	0	65,6%

Extraction des mots polarisés relatifs au domaine



Corpus	Score de classification			
	General Inquirer	Vocabulaire généraliste	Vocabulaire contextualisé	Vocabulaire spécialisé
SENT_100	57.5%	64.6%	65.6%	91.9%

Expérimentations sur d'autres jeux de données

A partir d'approches classiques par apprentissage supervisé

Method	SENT_100	DEFTPolitic	DEFTVideoGames	DEFTMovies
Naive bayes	51,5%	78,6%	90,5%	88,9%

A partir des lexiques d'opinions

Lexicons	Correct Scores		
	DEFTPolitic	DEFTVideoGames	DEFTMovies
GeneralInquirer	53,6%	65,5%	67,4%
LIWC	52,0%	54,4%	64,2%
JeuxDeMots	48,6%	50,0%	62,2%

A partir de l'approche OPITER

Corpus	Correct Score		
	GVPO	CVPO	SVPO
DEFTPolitic	54,0%	55,0%	69,8%
DEFTVideoGames	67,3%	68,4%	73,7%
DEFTMovies	77,6%	78,8%	82,6%

Conclusion

Travaux réalisés :

- Expérimentation d'approches classiques pour la classification de corpus relatifs à l'aménagement d'un territoire
- Méthode OPITER semi-automatisée pour la construction d'un vocabulaire d'opinions spécialisé relatif à l'aménagement d'un

Difficultés :

- Ressources françaises
- Complexité (hétérogénéité) des données territoriales

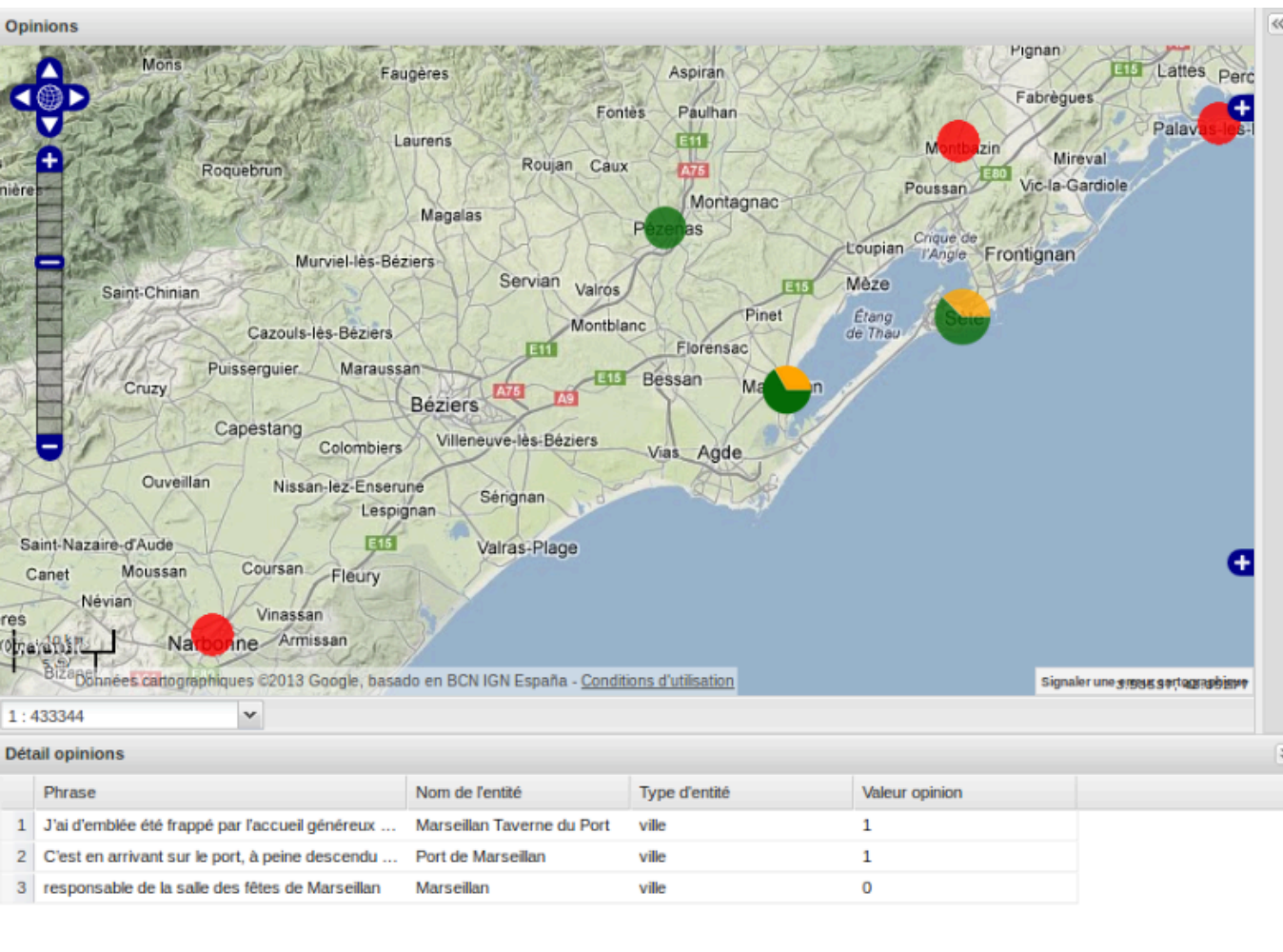
Travaux en cours :

- Identifier et qualifier les liens entre entités spatiales et opinions
- Identifier l'intensité des opinions

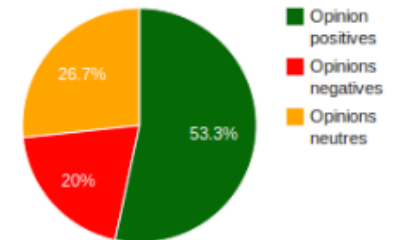


Eric Kergosien

Application to visualize opinions about geospatial entities on a map



Composante opinion



References

- Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics - COLING '04, pages 1367es, Morristown, NJ, USA, August 2004. for Computational Linguistics.
- Pak, A. (2012). Automatic, Adaptive, and Applicative Sentiment Analysis. University of Paris-Sud. Retrieved from <http://hal.upmc.fr/tel-00717329/>
- Piolat, A., Booth, R. J., Chung, C. K., Davids, M., & Pennebaker, J. W. (2011). La version française du LIWC : modalités de construction et exemples d'application. *Psychologie française*, 56, 145-159.
- Bestgen, N. V. Y. (n.d.). Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée. 2011.
- Gizem Gezici, Berrin Yanikoglu, Dilek Tapucu, Y. S. New Features for Sentiment Analysis: Do Sentences Matter? Proceedings of the 1st International Workshop on Sentiment Discovery from Affective Data (SDAD 2012), In conjunction with ECML-PKDD 2012
- Meng, X., Wei, F., Liu, X., Zhou, M., Li, S., & Wang, H. Entity-centric topic-oriented opinion summarization in twitter. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 2012
- Asiaee T., A., Tepper, M., Banerjee, A., & Sapiro, G. If you are happy and you know it... tweet. Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM 2012 (p. 1602). New York
- Kravchenko, A. Mining for Opinions Across Domains: A Cross-Language Study. Proceedings of the 1st International Workshop on Sentiment Discovery from Affective Data (SDAD 2012), In conjunction with ECML-PKDD 2012, Bristol
- Liu, B. (2012). Sentiment Analysis and Opinion Mining (p. 167). Morgan & Claypool Publishers.