Text Mining in Social Media

Objetivo

- 2 Casos (Genero / Variedad) --> desarrollados ambos

- A partir de un corpus de tweets preclasificados, modelo supervisado de clasificación: binomial/gender y multi/variety.

- Medidad de bondad de ajuste accuracy individual y conjunto (joint).

Estrategia de análisis

- Bolsa de palabras, n-gramas , nuevas variables
- Algoritmos de predicción : svm, ranger, reglog y kNN
- Creación de nuevas variables (gender)

Resultados

n	Model	CV	GENDER	VARIETY	JOINT
10	SVM	1	0,5257	0,1957	0,0992
50	SVM	1	0,6657	0,3764	0,2521
100	SVM	1	0,6778	0,4893	0,3257
500	SVM	1	0,7093	0,7578	0,5393
1000	SVM	1	0,659	0,782	
100	Ranger	1	0,6607	0,4979	0,3314
1000	Ranger	1	0,7164	0,885	0,629
100	RegLog	1	0,673		
500	RegLog	1	0,706		

Añadimos vble total palabras					
1000	Ranger		0,7193		

Añadimos distancias a min(gender), max(gender) y avg(gender)						
1000	Ranger		0,7229			
nuevas	Ranger		0,484			

Conclusiones

- Para gender :
  - Incrementar la bolsa de palabras aumenta ligeramente el accuracy.
  - El mejor resultado con: bolsa 1000 + nuevas variables.
  - Sin bolsa de plabrar empeora notablemente la predicción
  - Mejor modelo : bolsa 1000 + nuevas variables + Ranger → 0,723
- Para variety:
  - Incrementar la bolsa de palabras aumenta significativamente el accuracy.
  - Mejor modelo : bolsa 1000 + Ranger → 0,885

- El mejor algoritmo “Ranger” y el peor kNN.

Temas pendientes

- Nuevas variables para variety.
- N-gramas (no ha sido posible ne r probar con Python)
- Ensamble models (combinar mejores predicciones de diferentes algoritmos).