Natural Language Processing

In sentiment analysis, an accuracy of 80% is considered the to-be-expected value

Words are usually considered into groups of 2, 3 or even more words (bigrams, trigrams, n-grams) to account for the presence of senteces like "very funny" and "not very funny). Also, in a dataset there is a removal of stopwords (words without noticeable meaning)

If words are rapresented as one-hot vectors (every word is a vector of the size of the dictionary of words and only one bit is set to "1"), they are all equidistant. So we introduce word vectors, using the fascinating work of word2vec, a program that uses a neural network to predict relationship between words (example: "Athens is to Greece what Oslo is to...?" or "walking is to walked what swim is to...?). However, when using this model, we don't use one vector per word, but we use one vector per document which is the result of the average of the single word vectors. Unfortunately, this gives a worse performance than the one-hot encoded vectors.

recall: how many positives reviews detected divided by how many positive reviews are there.

recall: how many positives reviews correctly detected divided by how many positive reviews detected.

Ricordati che la precisione intorno ai 70% è ottima, in quanto ci sono tre bucket; se ce ne fossero solo due, allora dovresti avere

La stratiefied KFold cross validation è una tecnica che genera vari folds, ognungo con percentule più o meno uguale delle varie categorie di dato (cioè non c'è un fold con tutti commenti negativi).

La tecnica OneVsRestClassifier crea N classifier dove N è il numero di classi possibili dei dati. La tecnica OneVsOneClassifier crea N(N-1)/2 classifier, uno per ogni coppia di classi. Non si riscontrano differenze notevoli nella precisione del fit.

Bayes: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

Naive Bayes (si usano i log per evitare underflow e maggiore velocità) : $\hat{c} = \max_{c \in C} \left[log P(c) + \sum_i log P(w_i | c) \right]$ L'assunzione "naive" che si fa nel Naive Bayes è che tutti gli eventi w_i siano indipendenti, e quindi $P(w_0 \dots w_i | c) = P(w_0 | c) \dots P(w_i | c)$.

Inoltre, si assume che la posizione delle parole non conti; conta solo la loro frequenza!

La precision e la recall possono essere combinate con $F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$. Se $\beta < 1$, si da più importanza al recall; altrimenti si da più importanza alla precision. Se $\beta = 1$, entrambe hanno uguale importanza: $F_1 = \frac{2PR}{P+R}$. Quello che F è in realtà è la media armonica di precision e recall.

I classificatori generativi creano un modello per ogni classe e vedono per un documento quale modello lo rappresenta meglio. I classificatori discriminativi vedono quali feature sono importanti per distinguere le varie classi direttamente dall'input.

I lexicons sono dei dizionari disponibili online con delle parole già classificate come positive o negative.

chi_squared test: $X^2 = \sum \frac{(obs - exp)^2}{exp}$. Un chi2 molto alto significa che l'ipotesi nulla è falsa. Nel nostro caso, l'ipotesi nulla è che una parola non è collegata con la frequenza con cui sta in una classe.

L'altro collaboratore si chiama Antonio Norelli, fa il quinto anno della magistrale, e fa il secondo anno di magistrale a fisica. E' uno studente della ssas. Il suo numero di matricola è 1612487. Il suo username su slack (e su youtube) è noranta4.

Appunti

						. 1 1 1		
Qui ve	erranno listat	te varie pro	ove fatte;					
cv	Negative 1	Neutral P	ositive	Media				
3	.6706	.6536	.7271	.6837				
10	.6886	.6638	.7487	.7003				
20	.6851			.6997				
$\frac{1}{40}$.6883			.7020				
Folds		Neutral	Positive					
2	.6540	.6391	.7154	.6695				
3	.6689	.6517	.7183	.6797				
4	.6744	.6591	.7334	.6890				
5	.6782	.6552	.7386	.6907				
6	.6782	.6591	.7395	.6923				
7	.6805	.6522	.7414	.6914				
8	.6864	.6649	.7443	.6985				
9	.6812	.6608	.7401	.6940				
10	.6854	.6611	.7410	.6958				
11	.6863	.6587	.7406	.6952				
12	.6867	.6644	.7540	.7017				
13	.6865	.6654	.7508	.7009				
14	.6865	.6657	.7472	.6998				
15	.6844	.6661	.7449	.6985				
16	.6912	.6683	.7467	.7021				
17		.6649						
	.6833		.7477	.6986				
18	.6896	.6673	.7506	.7025				
19	.6876	.6663	.7428	.6989				
20	.6908	.6675	.7500	.7028				
21	.6894	.6708	.7435	.7012				
22	.6847	.6681	.7489	.7006				
23	.6865	.6673	.7462	.7000				
24	.6837	.6665	.7495	.6999				
25	6000	.6665	7504	7017				
20	.6883	.0000	.7504	.7017				
					NegativeR	NeutralR	PositiveR	MediaR
\mathbf{C}	NegativeP	NeutralP	Positive	eP MediaP	NegativeR .5812	NeutralR .7631	PositiveR .6216	MediaR .6553
C .01	NegativeP .6573	NeutralP .6087	Positive .7217	eP MediaP .6626	.5812	.7631	.6216	.6553
C .01 .02	NegativeP .6573 .6623	NeutralP .6087 .6225	Positive .7217 .7313	eP MediaP .6626 .6720	.5812 .5953	.7631 .7728	.6216 $.6281$.6553 $.6654$
C .01 .02 .03	NegativeP .6573 .6623 .6600	NeutralP .6087 .6225 .6354	Positive .7217 .7313 .7253	eP MediaP .6626 .6720 .6736	.5812 .5953 .6046	.7631 .7728 .7651	.6216 .6281 .6378	.6553 .6654 .6692
C .01 .02 .03 .04	NegativeP .6573 .6623 .6600 .6726	NeutralP .6087 .6225 .6354 .6396	Positive .7217 .7313 .7253 .7249	eP MediaP .6626 .6720 .6736 .6790	.5812 .5953 .6046 .6127	.7631 .7728 .7651 .7639	.6216 .6281 .6378 .6475	.6553 .6654 .6692 .6747
C .01 .02 .03 .04 .05	NegativeP .6573 .6623 .6600 .6726 .6698	NeutralP .6087 .6225 .6354 .6396 .6420	Positive .7217 .7313 .7253 .7249 .7270	eP MediaP .6626 .6720 .6736 .6790 .6796	.5812 .5953 .6046 .6127 .6103	.7631 .7728 .7651 .7639 .7659	.6216 .6281 .6378 .6475 .6503	.6553 .6654 .6692 .6747 .6755
C .01 .02 .03 .04 .05	NegativeP .6573 .6623 .6600 .6726 .6698 .6597	NeutralP .6087 .6225 .6354 .6396 .6420 .6499	Positive .7217 .7313 .7253 .7249 .7270 .7335	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810	.5812 .5953 .6046 .6127 .6103 .6257	.7631 .7728 .7651 .7639 .7659	.6216 .6281 .6378 .6475 .6503 .6443	.6553 .6654 .6692 .6747 .6755
C .01 .02 .03 .04 .05 .06	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662	NeutralP .6087 .6225 .6354 .6396 .6420 .6499	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793	.5812 .5953 .6046 .6127 .6103 .6257 .6236	.7631 .7728 .7651 .7639 .7659 .7615	.6216 .6281 .6378 .6475 .6503 .6443	.6553 .6654 .6692 .6747 .6755 .6771
C .01 .02 .03 .04 .05 .06 .07	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358	.7631 .7728 .7651 .7639 .7659 .7615 .7566	.6216 .6281 .6378 .6475 .6503 .6443 .6495	.6553 .6654 .6692 .6747 .6755 .6771 .6766
C .01 .02 .03 .04 .05 .06 .07 .08	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6559 .6530 .6601	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328 .7270	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837 .6897	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7635 .7558	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6559 .6530 .6601	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837 .6897	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328 .7270	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837 .6897 .6823	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7635 .7558	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764 .6675	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328 .7270 .7272	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837 .6897 .6823 .6850 .6904	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7635 .7558	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764 .6675 .6698	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579 .6617	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328 .7270 .7272 .7358	eP MediaP	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309 .6301	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7538 .7558	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479 .6604	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791 .6824 .6878
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16 .17 .18	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764 .6675 .6698 .6736 .6639	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579 .6617 .6526	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328 .7270 .7272 .7358 .7258 .7258	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837 .6897 .6823 .6850 .6904 .6808	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309 .6301 .6313	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7538 .7558 .7667 .7477 .7603	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479 .6604 .6665 .6544	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791 .6824 .6878 .6778
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16 .17 .18 .19	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764 .6675 .6698 .6736 .6639 .6754	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579 .6617 .6526 .6594 .6647	Positive	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837 .6897 .6823 .6850 .6904 .6808 .6909 .6954	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309 .6301 .6313 .6350 .6402	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7538 .7558 .7558 .7667 .7477 .7603 .7695	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479 .6604 .6665 .6544 .6689	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791 .6824 .6878 .6778 .6880 .6925
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16 .17 .18 .19 .2	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764 .6675 .6698 .6736 .6639 .6754 .6806	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579 .6617 .6526 .6594 .6647	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328 .7272 .7358 .7258 .7377 .7407 .7261	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837 .6837 .6897 .6823 .6850 .6904 .6808 .6909 .6954	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309 .6301 .6313 .6350 .6402	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7538 .7558 .7558 .7667 .7477 .7603 .7695 .7550	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479 .6604 .6665 .6544 .6689 .6677	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791 .6824 .6878 .6878 .6878 .6880 .6925 .6822
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16 .17 .18 .19 .2 .21	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764 .6675 .6698 .6736 .6639 .6736 .6639 .6754 .6806 .6760 .6848	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579 .6617 .6526 .6594 .6647 .6528	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328 .7270 .7272 .7358 .7258 .7377 .7407 .7261 .7317	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837 .6897 .6823 .6850 .6904 .6808 .6909 .6954 .6850 .6932	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309 .6301 .6313 .6350 .6402 .6293 .6394	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7538 .7558 .7558 .7558 .7667 .7477 .7603 .7695 .7550	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479 .6604 .6665 .6544 .6689 .6677 .6624	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791 .6824 .6878 .6878 .6925 .6822 .6906
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16 .17 .18 .19 .2 .21 .22	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764 .6675 .6698 .6736 .6639 .6736 .6639 .6754 .6806 .6760 .6848	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579 .6617 .6526 .6594 .6647 .6528 .6631 .6588	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328 .7270 .7272 .7358 .7258 .7261 .7317 .7288	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837 .6897 .6897 .6894 .6808 .6904 .6808 .6909 .6954 .6850 .6932 .6874	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309 .6301 .6313 .6350 .6402 .6293 .6394 .6269	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7538 .7558 .7558 .7667 .7477 .7603 .7695 .7550 .7611	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479 .6604 .6665 .6544 .6689 .6677 .6624 .6773	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791 .6824 .6878 .6778 .6880 .6925 .6822 .6906 .6848
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16 .17 .18 .19 .2 .21 .22 .23	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764 .6675 .6698 .6736 .6639 .6754 .6806 .6760 .6848 .6745 .6751	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579 .6617 .6526 .6594 .6647 .6528 .6631 .6588	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328 .7270 .7272 .7358 .7258 .7261 .7317 .7288 .7288 .7288	eP MediaP	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309 .6301 .6313 .6350 .6402 .6293 .6394 .6269 .6248	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7538 .7558 .7558 .7667 .7477 .7603 .7695 .7550 .7611 .7587 .7599	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479 .6604 .6665 .6544 .6689 .6677 .6624 .6713 .6689 .6640	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791 .6824 .6878 .6778 .6880 .6925 .6822 .6906 .6848 .6829
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16 .17 .18 .19 .2 .21 .22 .23 .24	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764 .6675 .6698 .6736 .6639 .6754 .6806 .6760 .6848 .6745 .6751 .6626	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579 .6617 .6526 .6594 .6647 .6528 .6631 .6588 .6536	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328 .7270 .7272 .7358 .7258 .7377 .7407 .7261 .7317 .7288 .7288 .7318	eP MediaP	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309 .6301 .6313 .6350 .6402 .6293 .6394 .6269 .6248	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7538 .7558 .7558 .7667 .7477 .7603 .7695 .7550 .7611 .7587 .7599 .7590	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479 .6604 .6665 .6544 .6689 .6677 .6624 .6713 .6689 .6640 .6657	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791 .6824 .6878 .6778 .6880 .6925 .6822 .6906 .6848 .6829 .6837
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16 .17 .18 .19 .2 .21 .22 .23 .24 .25	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764 .6675 .6698 .6736 .6639 .6754 .6806 .6760 .6848 .6745 .6751 .6626	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579 .6617 .6526 .6594 .6647 .6528 .6631 .6588 .6536	Positive .7217 .7313 .7253 .7249 .7270 .7335 .7215 .7344 .7295 .7242 .7247 .7308 .7328 .7270 .7272 .7358 .7258 .7278 .7407 .7261 .7317 .7288 .7288 .7318 .7283	eP MediaP	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309 .6301 .6313 .6350 .6402 .6293 .6269 .6248 .6265 .6390	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7538 .7558 .7558 .7667 .7477 .7603 .7695 .7550 .7611 .7587 .7599 .7590 .7498	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479 .6604 .6665 .6544 .6689 .6677 .6624 .6713 .6689 .6640 .6657 .6665	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791 .6824 .6878 .6778 .6880 .6925 .6822 .6906 .6848 .6829 .6837 .6851
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16 .17 .18 .19 .2 .21 .22 .23 .24 .25 .26	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764 .6675 .6698 .6736 .6639 .6754 .6806 .6760 .6848 .6745 .6751 .6626 .6753 .6675	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579 .6617 .6526 .6594 .6647 .6528 .6631 .6588 .6536 .6630 .6587	Positive	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837 .6897 .6823 .6850 .6904 .6808 .6909 .6954 .6850 .6932 .6874 .6858 .6858	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309 .6301 .6313 .6350 .6402 .6293 .6394 .6269 .6248 .6265 .6390 .6245	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7538 .7558 .7558 .7667 .7477 .7603 .7695 .7550 .7611 .7587 .7599 .7590 .7498 .7570	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479 .6604 .6665 .6544 .6689 .6677 .6624 .6713 .6689 .6640 .6657 .6665 .6665	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791 .6824 .6878 .6878 .6878 .6925 .6822 .6906 .6848 .6829 .6837 .6851 .6825
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16 .17 .18 .19 .2 .21 .22 .23 .24 .25 .26 .27	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6675 .6698 .6736 .6674 .6806 .6754 .6806 .6760 .6848 .6745 .6751 .6626 .6753 .6675	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579 .6617 .6526 .6594 .6647 .6528 .6631 .6588 .6536 .6630 .6587 .6570 .6643	Positive	eP MediaP .6626 .6720 .6736 .6790 .6796 .6810 .6793 .6869 .6860 .6817 .6827 .6837 .6897 .6823 .6850 .6904 .6808 .6909 .6954 .6850 .6932 .6874 .6858 .6858 .6874 .6858	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309 .6301 .6313 .6350 .6402 .6293 .6394 .6269 .6248 .6265 .6390 .6245 .6402	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7534 .7538 .7558 .7558 .7558 .7667 .7477 .7603 .7695 .7550 .7611 .7587 .7599 .7590 .7498 .7570 .7603	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479 .6604 .6665 .6544 .6689 .6677 .6624 .6713 .6689 .6640 .6657 .6665 .6661 .6746	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791 .6824 .6878 .6778 .6880 .6925 .6822 .6906 .6848 .6829 .6837 .6851 .6825 .6917
C .01 .02 .03 .04 .05 .06 .07 .08 .09 .1 .11 .12 .13 .14 .15 .16 .17 .18 .19 .2 .21 .22 .23 .24 .25 .26	NegativeP .6573 .6623 .6600 .6726 .6698 .6597 .6662 .6692 .6695 .6636 .6674 .6673 .6764 .6675 .6698 .6736 .6639 .6754 .6806 .6760 .6848 .6745 .6751 .6626 .6753 .6675	NeutralP .6087 .6225 .6354 .6396 .6420 .6499 .6503 .6570 .6591 .6574 .6559 .6530 .6601 .6524 .6579 .6617 .6526 .6594 .6647 .6528 .6631 .6588 .6536 .6630 .6587	Positive	eP MediaP	.5812 .5953 .6046 .6127 .6103 .6257 .6236 .6358 .6257 .6188 .6253 .6265 .6301 .6337 .6309 .6301 .6313 .6350 .6402 .6293 .6394 .6269 .6248 .6265 .6390 .6245	.7631 .7728 .7651 .7639 .7659 .7615 .7566 .7631 .7639 .7679 .7534 .7534 .7538 .7558 .7558 .7667 .7477 .7603 .7695 .7550 .7611 .7587 .7599 .7590 .7498 .7570	.6216 .6281 .6378 .6475 .6503 .6443 .6495 .6511 .6608 .6511 .6620 .6624 .6677 .6479 .6604 .6665 .6544 .6689 .6677 .6624 .6713 .6689 .6640 .6657 .6665 .6665	.6553 .6654 .6692 .6747 .6755 .6771 .6766 .6833 .6835 .6793 .6802 .6808 .6871 .6791 .6824 .6878 .6878 .6880 .6925 .6822 .6906 .6848 .6829 .6837 .6851 .6825

Beta	Negative			ositive		NegativeR		PositiveR	MediaR
.5	.6623	.661		.7292	.6842	.6394	.7437	.6628	.6820
.51	.6662	.659		.7323	.6860	.6313	.7550	.6641	.6835
.52	.6682	.657		.7324	.6862	.6422	.7461	.6620	.6835
.53	.6682	.654		.7298	.6843	.6273	.7538	.6641	.6817
.54	.6612	.652		.7191	.6777	.6265	.7469	.6527	.6754
.55	.6655 .6626	.650 .653		.7318 .7315	.6826 .6826	.6245 $.6325$.7514 $.7473$.6632	.6797
.56		.653		.7262	.6826 .6815	.6325 .6329	.7473 .7493	.6596	.6798
.57 .58	.6645 $.6724$.659		.7275	.6865	.6329	.7493 .7591	.6544 $.6592$.6789 $.6837$
.50 .59	.6614	.658		.7337	.6845	.6252	.7619	.6584	.6818
.6	.6768	.658		.7314	.6888	.6277	.7615	.6693	.6862
.61	.6705	.650		.7250	.6819	.6245	.7570	.6548	.6787
.62	.6654	.658		.7245	.6830	.6341	.7522	.6560	.6808
.63	.6680	.654		.7324	.6849	.6305	.7554	.6600	.6820
.64	.6631	.654		.7357	.6844	.6220	.7619	.6604	.6814
.65	.6677	.653		.7218	.6811	.6313	.7514	.6527	.6785
.66	.6668	.657		.7351	.6864	.6293	.7627	.6564	.6828
.67	.6587	.645		.7309	.6783	.6253	.7465	.6527	.6748
.68	.6692	.657		.7320	.6863	.6321	.7599	.6584	.6835
.69	.6780	.651		.7293	.6862	.6192	.7639	.6661	.6831
.7	.6684	.652		.7299	.6837	.6261	.7582	.6584	.6809
.71	.6674	.648		.7250	.6801	.6285	.7526	.6499	.6770
.72	.6670	.646		.7266	.6801	.6204	.7566	.6535	.6769
.73	.6675	.650)1	.7262	.6813	.6200	.7566	.6588	.6785
.74	.6672	.648	32	.7231	.6795	.6200	.7566	.6519	.6762
.75	.6736	.650)3	.7275	.6838	.6350	.7570	.6495	.6805
.76	.6658	.649	95	.7291	.6814	.6196	.7566	.6596	.6786
.77	.6678	.652	20	.7299	.6832	.6297	.7595	.6511	.6801
.78	.6733	.652	22	.7299	.6851	.6192	.7611	.6661	.6821
.79	.6762	.655	50	.7257	.6857	.6216	.7590	.6685	.6830
.8	.6740	.656	30	.7353	.6884	.6337	.7667	.6540	.6848
.81	.6631	.645	55	.7297	.6794	.6176	.7619	.6475	.6756
.82	.6721	.650)2	.7318	.6847	.6228	.7655	.6552	.6812
.83	.6699	.651		.7267	.6828	.6156	.7655	.6580	.6797
.84	.6724	.653		.7270	.6843	.6321	.7570	.6552	.6814
.85	.6619	.649		.7276	.6796	.6232	.7558	.6499	.6763
.86	.6683	.643		.7275	.6798	.6119	.7586	.6576	.6760
.87	.6762	.642		.7443	.6878	.6305	.7675	.6499	.6826
.88	.6779	.647		.7276	.6844	.6180	.7683	.6564	.6809
.89	.6635	.651		.7277	.6808	.6139	.7671	.6519	.6777
.9	.6814	.658		.7297	.6898	.6325	.7700	.6580	.6868
.91	.6699	.652		.7289	.6839	.6232	.7639	.6552	.6808
.92	.6741	.651		.7202	.6818	.6204	.7594	.6572	.6790
.93	.6700	.650		.7305	.6836	.6309	.7607	.6479	.6798
.94	.6714	.646		.7324	.6835	.6249	.7594	.6548	.6797
.95	.6690	.647		.7362	.6843	.6261	.7643	.6507	.6804
.96	.6712	.644		.7255	.6804	.6151	.7582	.6576	.6770
.97	.6772	.646		.7313	.6851	.6224	.7639	.6580	.6814
.98 .99	.6731 .6700	.649 .649		.7333 .7315	.6831 .6836	.6184 .6180	.7667 .7691	.6503 $.6531$.6785 $.6801$
						.0100	.7091	.0551	.0001
	n-grams 4	analyzer char	binary	$\frac{\text{cv}}{3}$	average f1 .6869				
	6	char	true		.6990				
	8	char	true	$\frac{3}{3}$					
	8	char	true true	3 12	.6991 $.7104$				
		analyzer	binary		average f1				
	n-grams 4	char	false	$\frac{\text{cv}}{3}$.6926				
	8	char	false	3	.7004				
	6	char	false	12	.7101				
	fatte a n		_0.250						

Prove fatte a mano:

Prendendo 100 commenti (quasi) a caso, classificandoli per conto mio ottengo 96/100 di correttezza.

Prendendo 50 commenti da quelli sbagliati, ottengo il 37/50 di correttezza.

Usando le lettere al posto delle parole, si arriva a guadagnare un punto e mezzo di percentuale di f1. (Tuttavia ancora da provare con binary a True)

Se si usano tanti n-gram, l'impatto del random è un pochino più basso.

Il Tfidf fa schifo, lascia perdere. Molto meglio avere le occorrenze binarie delle parole.

IDEE

Giocare molto, **molto** con i possibili argomenti di count Vectorizer. Ad esempio min_df e max_df potrebbero essere utili. La cosa che ha detto Chierichetti: non si può avere un train data fatto bene; bisogna assegnare un punteggio ai commenti, non una categoria.

Devo proprio mettere shuffle=False in classifier_report()?

LE COSE NON SHUFFLATE FANNO CAGARE RISPETTO ALLE COSE SHUFFLATE.

Provare word vectors invece che bag-of-words

Provare Keras.

Cambiando parametri del machine learning cambia veramente poco. Molto di più il countVectorizer.

Il countVectorizer è molto lento; invece la logistic regression no. Perchè?

Qual è la parte dell'algortimo che usa multithreading e quale no?

Perchè fate un estimator fatto da voi invece che usare normalmente la logistic regression?

Domande:

I commenti che l'algoritmo trova sbagliati sono spesso e volentieri sbagliati. Perchè?

Spesso capita che le probabilità sono molto vicine tra loro e l'algoritmo sbaglia poco.

Le lettere in aggiunta potrebbero essere parole sbagliate.

Perchè i dati sono una sparse matrix?

Perchè la sparse matrix va scalata?

Non si potrebbero fare classificatori diversi per ogni sito?

Matrix correlation?? Glove dataset.

Perchè avete filtrato via gli hashtags?

Posso toccare la funzione build_dataset()? E Tagger?

Ho parlato con Flavio Chierichetti.

Ciao gabriele,

PUNTATINA?