

Verslag Practicum 2

Martijn Koenis (3770214) Jordi Vermeulen (3835634)

Samenvatting

1 Probleem en omschrijving corpus

We hebben het volgende classificatieprobleem onderzocht: het indelen van songteksten in de genres Blues, Country, Folk, Gospel, Metal, R&B, Rap en Soul. Om dat te kunnen doen hebben we een corpus verzameld van 50 songteksten per genre (in totaal 400). Deze songteksten hebben we op internet gezocht. Ter inspiratie voor de keuze van nummer hebben we indien nodig gebruik gemaakt van hitlijsten en verzamel-CD's. We hebben ervoor gezorgd dat we geen dubbele nummers hebben door een script te schrijven die sterk op elkaar lijkende txt-bestanden detecteert en rapporteert.

2 Voorbewerking

Voordat we de teksten gaan analyseren halen we eerst alle whitespace weg zodat die niet interfereren met de uitkomst. Vervolgens vervangen we alle hoofdletters door normale letters omdat die voor ons doeleinde hetzelfde betekenen. Daarna verwijderen we alle Engelse stopwoorden omdat deze in alle genres voorkomen en dus niet interessant zijn om te bekijken. Ook verwijderen we alle getallen in de tekst. Voor de classificatie hebben we het k nearest neighbors, support vector machine en naïve bayes algoritme gebruikt. Voor alle drie de algoritmen hebben we getest welke soort document term matrix het beste werkte. We hebben biray, raw en tfidf document term matrices geprobeerd. Voor alle drie de algoritme kwam de binary matrix als beste naar boven en deze hebben we dus gebruikt.

3 Parameters

Voor alle parameters hebben we uitgebreid getest wat de beste waarden was. Zowel met behulp van de tune functie als handmatige tests. Voor het knn algoritme gebruiken we een k van 9. Voor het svm algoritme gebruiken we een cost van 10 en een gamma van 0.001.

4 Data

We hebben voor elk algoritme 2 verschillende soorten tests uitgevoerd. Ten eerste hebben we voor gekeken naar de werking op de hele set van genres. De

Tabel 1: NB, 25, Pairs

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	0.0	73.6	88.8	95.6	92.4	95.6	100.0	80.4
Country	66.8	0.0	75.6	86.4	90.0	92.4	100.0	70.8
Folk	56.4	58.0	0.0	68.4	61.2	88.8	99.6	82.8
Gospel	85.6	74.0	82.4	0.0	61.2	86.8	97.2	80.4
Metal	89.6	81.2	84.8	74.4	0.0	88.8	98.0	84.4
R&B	69.6	60.8	73.2	73.6	71.6	0.0	80.0	58.0
Rap	90.8	88.0	91.2	88.8	88.4	81.2	0.0	90.4

Tabel 2: NB, 25, All

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	56.8	25.2	13.6	2.8	2.4	2.8	2.8	18.0
Country	14.4	22.0	10.8	4.0	2.4	8.8	3.2	20.8
Folk	12.8	12.8	30.0	13.2	18.4	4.0	0.0	4.0
Gospel	1.6	8.8	9.6	36.8	19.6	6.8	0.4	4.4
Metal	2.4	12.0	24.0	32.8	49.2	4.8	0.4	7.2
R&B	4.0	3.2	7.6	6.4	4.0	35.2	13.2	19.2
Rap	0.0	0.0	0.0	0.0	0.8	21.6	74.0	0.4

data van deze tests is te zien in de even tabellen. In de cel (Country, Folk) staat het percentage van de Country nummers die als Folk zijn gelabeld door het algoritme. Daarnaast hebben we gekeken naar de werking op alle subsets van 2 genres. De data daarvan is te vinden in de oneven tabellen. In de cel (Country, Folk) staat het percentage van de Country nummers dat goed wordt geclassificeerd als de genres Country en Folk worden getest. Voor iedere soort test hebben we ook gekeken naar 3 verschillende groottes van de trainingsset namelijk 10, 25 en 40. De grote van de testset staat in de naam van de tabel samen met het algoritme en het soort test ("All" voor alle genres en "Pairs" voor de subset test).

A description of the training and testing of the classifiers, and a comparison of their predictive performance. Interpretation of the results. Why did you get good or bad prediction results? Why do you think did one algorithm or weighting scheme work better than another?

Average 41.25
Average 36.8125
Average 42.125
Average 25.3
Average 22.875
Average 27
Average 40.2
Average 34.25
Average 40.375

Tabel 3: NB, 10, Pairs

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	0.00	67.75	95.25	96.75	95.25	98.25	100.00	84.00
Country	60.75	0.00	83.00	86.50	81.75	96.25	100.00	68.50
Folk	38.25	33.75	0.00	52.75	62.50	92.00	100.00	41.00
Gospel	72.25	59.50	82.00	0.00	67.25	94.25	100.00	61.25
Metal	69.50	72.75	67.50	61.25	0.00	92.75	99.75	69.25
R&B	41.75	40.50	64.25	48.75	62.00	0.00	93.25	39.25
Rap	52.75	62.25	69.75	74.75	72.50	45.25	0.00	57.00

Tabel 4: NB, 10, All

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	57.50	27.25	23.25	9.00	6.50	9.75	3.75	23.25
Country	12.75	22.75	16.00	7.00	7.00	8.75	4.75	20.00
Folk	8.75	15.00	18.75	10.50	11.50	3.75	2.25	6.50
Gospel	3.00	9.75	21.25	43.25	29.00	10.50	0.50	8.75
Metal	3.00	8.75	14.25	18.50	38.25	6.25	1.75	5.50
R&B	2.00	2.00	1.50	4.75	1.50	26.50	18.75	9.25
Rap	0.00	0.00	0.25	0.25	0.50	15.25	60.75	0.00

Tabel 5: NB, 40, Pairs

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	0	66	81	95	96	90	100	75
Country	74	0	70	85	90	89	100	68
Folk	71	50	0	72	69	88	99	81
Gospel	81	82	76	0	70	83	97	84
Metal	95	83	75	80	0	87	98	87
R&B	75	69	73	72	83	0	79	64
Rap	94	93	89	94	86	76	0	88

Tabel 6: NB, 40, All

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	58	19	16	6	2	0	1	16
Country	16	21	10	3	1	4	2	18
Folk	7	16	32	13	16	3	0	4
Gospel	1	12	7	39	25	8	2	5
Metal	3	7	24	30	46	2	0	6
R&B	6	4	5	8	5	32	9	18
Rap	0	1	0	0	0	27	78	2

Tabel 7: KNN, 25, Pairs

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	0.0	53.2	59.6	89.2	80.8	99.6	100	86.0
Country	74.4	0.0	58.8	90.4	78.8	98.4	100	79.2
Folk	77.2	54.4	0.0	93.2	69.2	99.2	100	71.6
Gospel	76.8	43.2	20.4	0.0	54.8	91.6	100	78.0
Metal	85.2	62.8	54.0	66.0	0.0	99.6	100	88.8
R&B	26.0	27.6	32.0	54.0	39.2	0.0	100	24.4
Rap	22.4	21.2	23.2	34.0	27.6	16.4	0	14.4

Tabel 8: KNN, 25, All

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	24.0	4.8	6.0	2.8	0.8	8.0	12.0	16.8
Country	15.6	17.2	10.8	10.4	6.0	15.2	6.0	22.0
Folk	32.8	30.8	29.6	19.2	23.6	15.6	8.8	19.2
Gospel	9.2	12.0	10.4	22.4	17.6	11.2	4.0	9.2
Metal	15.6	28.0	41.2	39.2	49.6	15.6	5.2	22.0
R&B	0.0	0.8	0.0	2.8	0.0	13.2	13.2	1.6
Rap	0.0	0.0	0.0	0.0	0.0	7.6	37.2	0.0

Tabel 9: KNN, 10, Pairs

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	0.00	68.25	90.75	96.25	92.75	98.00	100	84.25
Country	45.50	0.00	70.25	79.25	84.50	97.00	100	64.50
Folk	16.00	35.50	0.00	60.50	31.25	77.50	100	36.75
Gospel	15.25	42.00	53.50	0.00	30.50	89.50	100	61.25
Metal	29.25	35.75	76.50	82.00	0.00	97.75	100	76.50
R&B	12.25	23.00	47.75	36.00	20.75	0.00	100	12.50
Rap	0.25	3.50	8.75	2.25	4.75	2.25	0	0.25

Tabel 10: KNN, 10, All

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	28.00	16.25	12.00	4.00	4.00	16.00	18.25	22.00
Country	14.25	15.25	10.50	7.75	6.00	12.50	6.00	11.25
Folk	18.25	18.50	19.75	14.50	14.50	6.75	5.00	8.25
Gospel	11.25	12.75	17.25	25.50	24.75	8.50	6.75	15.50
Metal	23.25	29.75	36.50	42.00	46.75	22.00	9.75	26.00
R&B	0.75	1.50	0.50	1.50	0.50	12.00	15.75	3.00
Rap	0.00	0.00	0.00	0.00	0.00	2.25	21.75	0.00

Tabel 11: KNN, 40, Pairs

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	0	58	47	92	83	100	100	90
Country	84	0	54	99	65	100	100	78
Folk	84	64	0	93	81	100	100	90
Gospel	79	43	22	0	37	95	100	83
Metal	94	76	32	79	0	98	100	91
R&B	41	28	45	50	50	0	99	19
Rap	35	31	35	43	38	21	0	28

Tabel 12: KNN, 40, All

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	22	9	4	2	0	1	7	12
Country	11	15	4	14	7	13	7	26
Folk	37	32	38	25	26	11	5	12
Gospel	6	5	9	13	12	12	3	8
Metal	19	30	45	44	52	15	7	26
R&B	0	0	0	0	0	19	8	1
Rap	0	0	0	0	0	8	42	0
Soul	5	9	0	2	3	21	21	15

Tabel 13: SVM, 25, Pairs

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	0.0	69.2	78.0	92.8	96.0	86.8	100.0	70.0
Country	67.6	0.0	75.2	82.0	89.6	86.0	100.0	66.0
Folk	76.0	51.6	0.0	78.8	69.6	88.0	96.8	76.0
Gospel	80.8	80.8	77.2	0.0	64.0	78.8	94.8	80.8
Metal	83.2	83.6	75.2	71.2	0.0	81.2	98.8	82.8
R&B	72.4	65.2	80.0	69.6	77.6	0.0	84.8	56.8
Rap	90.4	86.4	92.0	90.0	93.2	68.8	0.0	86.4

Tabel 14: SVM, 25, All

	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	51.2	20.0	14.8	3.2	3.2	4.4	2.4	12.0
Country	20.0	32.4	18.8	4.8	6.4	11.6	4.4	26.8
Folk	12.0	12.4	26.4	12.8	18.4	2.8	1.2	6.0
Gospel	5.2	6.4	10.8	45.2	23.2	9.2	0.0	5.2
Metal	1.2	9.2	20.8	25.6	39.6	7.2	1.2	4.8
R&B	0.8	7.6	3.6	5.2	4.4	32.8	24.4	14.0
Rap	0.4	0.8	0.4	0.0	0.0	14.8	62.8	0.0

Tabel 15: SVM, 10, Pairs								
	Blues	Country	Folk	Gospel	Metal	R&B	Rap	Soul
Blues	0.00	81.25	90.75	94.50	94.50	92.00	100.00	79.50
Country	45.25	0.00	74.00	90.00	88.75	91.75	100.00	60.25
Folk	43.75	48.75	0.00	68.50	60.00	86.00	100.00	58.75
Gospel	75.00	69.25	76.75	0.00	69.25	80.25	98.25	73.75
Metal	71.00	65.75	69.50	63.25	0.00	84.00	99.50	71.00
R&B	66.25	58.00	80.00	76.00	76.75	0.00	87.50	47.50
Rap	76.00	78.25	76.75	81.25	85.25	53.00	0.00	63.75

Tabel 16: SVM, 10, All