

Text Mining

Prof. Klaus-Robert Müller
Ulf Brefeld

Arbeitsgruppe Maschinelles Lernen



Why Text Mining?

- Textual information is ubiquitous.
 - WWW, news archives, linked document archives, ...
- Information extraction.
 - Relation and event extraction.
 - Find entities like names, date, time, location, ...
- Information retrieval.
 - Web search.
 - Find related (news) articles.
- Applications based on text mining:
 - Search engines (e.g., Yahoo, Google).
 - Recommender systems (e.g., Amazon).
 - Machine translation (e.g., babelfish).

Overview

- 1 Characteristics of Natural Language
- 2 Document Representations
- 3 Applications
- 4 Summary & Further Applications

Characteristics of Natural Language

Eigenschaften natürlicher Sprache

- Unendlich viele Ausdrücke.
- Rekursion:
 - *Der Bezug des Bettes des Hotels des Ermittlungsteams der Ursache des Absturzes des Systems ...*
 - *Systemabsturzursachenermittlungsteamhotelbettbezug*
- Konjunktion (Aufzählung):
 - *Am Sonntag fraß Sie sich durch einen Äpfel, zwei Bananen, drei Tomaten, vier Gurken, fünf Schokohasen, sechs ...*
- Hinzunahme neuer Basiselemente:
 - Entlehnung: *to go, Email, ...*
 - Kreativität: *unkaputtbar, Handy, ...*

Eigenschaften natürlicher Sprache (forts.)

- Synonyme: *zwölf*, *12* und *XIII* ; *Orange* und *Apfelsine*,...
- Homonyme: *Schloss* (Gebäude und Türschloss)
- Ambiguität: *Ich sehe den Mann mit dem Fernrohr, Staubecken*....

Desambiguierung

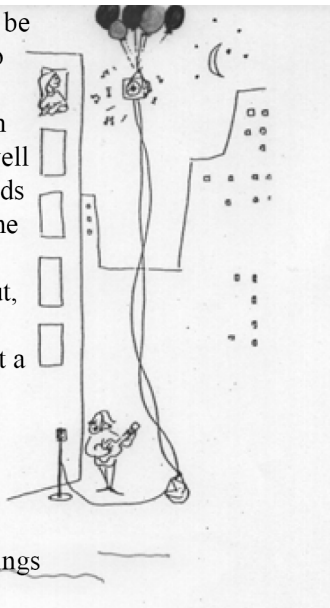
- Kontextabhängig.
- Beispiel: *Nach 14 Jahren Kohl, ...*
 - ... *wollten wir mal wieder etwas anderes essen.*
 - ... *lag die Arbeitslosigkeit bei x%.*
- Manchmal reicht das nicht...

Do you understand English?

If the balloons popped, the sound wouldn't be able to carry since everything would be too far away from the correct floor. A closed window would also prevent the sound from carrying, since most buildings tend to be well insulated. Since the whole operation depends on a steady flow of electricity, a break in the middle of the wire would also cause problems. Of course, the fellow could shout, but the human voice is not loud enough to carry that far. An additional problem is that a string could break on the instrument. Then there could be no accompaniment to the message. It is clear that the best situation would involve less distance. Then there would be fewer potential problems. With face to face contact, the least number of things could go wrong.

(Bransford and Johnson (1973))

If the balloons popped, the sound wouldn't be able to carry since everything would be too far away from the correct floor. A closed window would also prevent the sound from carrying, since most buildings tend to be well insulated. Since the whole operation depends on a steady flow of electricity, a break in the middle of the wire would also cause problems. Of course, the fellow could shout, but the human voice is not loud enough to carry that far. An additional problem is that a string could break on the instrument. Then there could be no accompaniment to the message. It is clear that the best situation would involve less distance. Then there would be fewer potential problems. With face to face contact, the least number of things could go wrong.



Common words in *Tom Sawyer*

word	frequency
the	3332
and	2972
a	1775
to	1725
of	1440
was	1161
it	1027
in	906
that	877
he	877
I	783
his	772
you	686
Tom	679

word frequency	freq. of frequency
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11-50	540
51-100	99
> 100	102

Zipf's Law

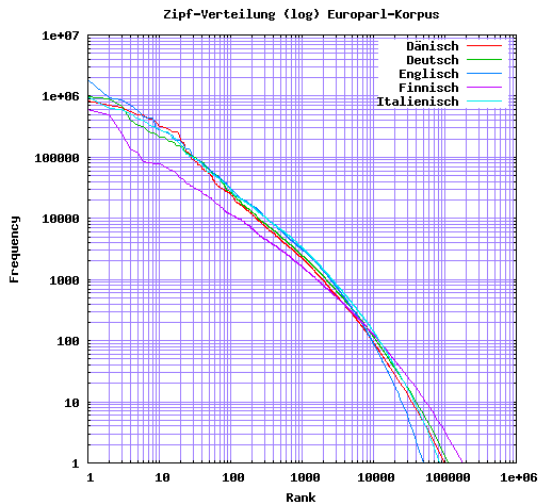
- Explores the relationship between the frequency of a word f and its rank r (i.e., its position in the list).

$$f \propto \frac{1}{r}$$

or in other words: There is a constant k such that $f \cdot r = k$.

- Example: the 50th most common word should occur with three times the frequency of the 150th most common word.
- Zipf distribution: A few very frequent words, a middling number of medium frequency words, and many uncommon words.

Exemplary Zipf Distribution



Empirical evaluation of Zipf's law on *Tom Sawyer*

word	freq.	rank	$f \cdot r$	word	freq.	rank	$f \cdot r$
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	Could	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

Document Representations

Tokenization

- Document = sequence of characters or symbols.
- Tokenization: Convert a document into a sequence of *tokens*.
- A token is a categorized block of text.
 - *Bello chases the cat.* → $\langle \text{Bello} \mid \text{chases} \mid \text{the} \mid \text{cat} \rangle$
- Frequently, prior knowledge is necessary:

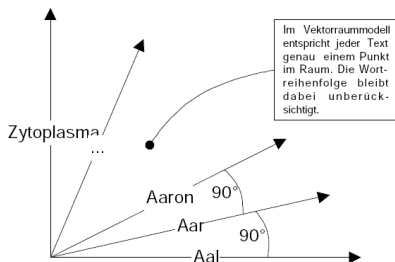
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000

ET QUOS MORTE FIXI EST ISTE LECTI
 PERCORPUS PAMAH
 CYSI FUSQSAE REBUS
 QUI EXMORTUIS RESURREXIT
 UT FRUCTIFICOSUS SED
 UMENTI ESSE MUSI INOXA
 PASSIONES PECATORUM
 QUAE PER LEGEM ANTO PERBANTUR
 IN MEMBRIS NOSTRIS
 UT FRUCTIFICARE IN MORTI
 NUNCIAMENS SOLUTIBUS
 ALECEMORTIS
 INQUA DETINEBAMUR
 ITA UT SE RUAMUS IN NOUIT RESP
 ET NON INIUGULATI LITTEAE
 QUAERERE DECEMUS
 LEX PECATORUM TABSIT
 SED PECCATUM NON COGNOU
 NIS PER LEGEM
 NAM CONCUSPICE NTIAM NESCIEBAM
 NIS ILEXICERET

漢書卷六十八
 霍光金日磾傳第三十八
 霍光字子真，安定郡高平人也。少為將，西征匈奴，入為右將軍，領兵，封博陸侯。光秉政，與金日磾同心。光薨，光妻毒殺金日磾，光妻毒殺金日磾，光妻毒殺金日磾。

٩٢ . سورة الليل بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِیْمِ وَاللَّیْلِ إِذَا
 یَغْشَى وَالنَّهَارِ إِذَا تَجَلَّى وَمَا خَلَقَ الذَّكْرَ وَالْأُنثَى إِنَّ
 سَعِیْكُمْ لَسُنَّی فَأَمَّا مَنْ أُعْطَى وَآتَى وَصَدَّقَ
 بِالْخُسْنَى فَسَنُیْسِرُهُ لِلْیُسْرَى وَأَمَّا مَنْ بَخِلَ
 وَاسْتَعْتَى وَكَذَّبَ بِالْخُسْنَى فَسَنُیْسِرُهُ لِلْعُسْرَى وَمَا
 یُعْیِبُ عَنْهُ مَالُهُ إِذَا تَرَدَّى إِنَّ عَلَیْنَا لِلْهَدَى وَإِنْ لَنَا
 لِلْآخِرَةِ وَالْأُولَى فَأَنْذَرْتُكُمْ نَارًا تَلَظَّى لَا یَصْلُهَا إِلَّا
 الْأَشْقَى الَّذِی كَذَّبَ وَتَوَلَّى وَسَیْجَنَّبُهَا الْأَتْقَى الَّذِی یُؤْتِی
 مَالَهُ یَتْرُكْی وَمَا لِأَحَدٍ عِنْدَهُ مِنْ نِعْمَةٍ تُجْزَى إِلَّا ابْتِغَاءَ
 وَجْهِ رَبِّهِ الْأَعْلَى وَلَسَوْفَ یَرْضَى

The Vector Space Model



- Documents are represented in a high-dimensional vector space.
- Axes are identified with tokens.
- Ordering of tokens is lost.
- Examples: Bag-of-words, TF.IDF representations.

Bag-of-Words Representation

- Let $D = \{d_1, \dots, d_m\}$ be a set of documents.
- Build dictionary $\mathcal{D} = \bigcup_{d \in D} \{w : \text{token } w \text{ occurs in document } d\}$.
- Indicator function $\mathbb{I}[z] = 1$ if z is true and 0 otherwise.

$$\text{BOW}(d_j) = \begin{pmatrix} \mathbb{I}[w_1 \in d_j] \\ \mathbb{I}[w_2 \in d_j] \\ \vdots \\ \mathbb{I}[w_{|\mathcal{D}|} \in d_j] \end{pmatrix}$$

- Drawback: All tokens are equally important.

TF.IDF

- Term frequency: number of occurrences of term w_i in a document.
- Problem 1: Long documents have large term frequencies
⇒ difficult for similarity measure.
- Solution: normalize term frequency.

$$TF(w_i) = \frac{TF(w_i)}{\sum_i TF(w_i)}$$

- Problem 2: Several words are irrelevant (e.g., *the*, *and*, ...)
- Solution: inverse document frequency.

$$IDF(w_i) = \frac{\# \text{ documents}}{\# \text{ documents containing } w_i}.$$

TF.IDF Representation

- TF.IDF representation of word w_i determined by $TF(w_i) \cdot IDF(w_i)$.
- TF.IDF representation of document d_j is given by

$$TF.IDF(d_j) = \begin{pmatrix} TF(w_1) \cdot IDF(w_1) \\ TF(w_2) \cdot IDF(w_2) \\ \vdots \\ TF(w_{|D|}) \cdot IDF(w_{|D|}) \end{pmatrix}$$

N-grams

- Ordering in BOW and TF.IDF representation is lost.
- BUT: neighboring tokens are not independent!
- N-grams represent sequences up to n tokens:

$$P(w_t | w_{t-n+1}, \dots, w_{t-1}) = \frac{P(w_{t-n+1}, \dots, w_t)}{P(w_{t-n+1}, \dots, w_{t-1})}$$

- Several n -gram representations are possible:
 - Occurrence: $NG(w_1, \dots, w_n; d_j) = \mathbb{I}((w_1, \dots, w_n) \in d_j)$
 - Frequency: $NG(w_1, \dots, w_n; d_j) = \#((w_1, \dots, w_n) \in d_j)$
 - Probabilistic: $NG(w_1, \dots, w_n; d_j) = P(w_n | w_1, \dots, w_{n-1}; d_j)$

N -gram Representations

- N -gram vector space has one dimension per n -gram.
- Let \mathcal{N} consist of all possible $(|\mathcal{D}|^n)$ n -grams.
- The n -gram representation of document d_j is given by

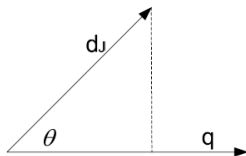
$$NGram(d_j) = \begin{pmatrix} NG(\mathbf{w}_1, d_j) \\ NG(\mathbf{w}_2, d_j) \\ \vdots \\ NG(\mathbf{w}_{|\mathcal{N}|}, d_j) \end{pmatrix}, \quad \mathbf{w} \in \mathcal{N}$$

- Parameter n needs to be chosen appropriately.

Normalization

- Problem: long texts result in long feature vectors.
 - Example: web search where queries hardly consist of more than 3 tokens.
- Solution: normalize feature vectors such that $\|\phi(d_j)\| = 1$ for all j .
- Similarity between document d_j and query q given by

$$\text{sim}(d_j, q) = \cos(d_j, q) = \frac{\langle d_j, q \rangle}{\|d_j\| \|q\|}$$



Dimensionality Reduction

- BOW, TF.IDF, and n -gram feature spaces are high-dimensional.
- Problems when using non-sparse learners (e.g., naïve Bayes).
- Stemming.
 - Strip off *affixes* (remove inflectional endings of words).
 - E.g., map for occurrences of *go*, *gone*, *going*, etc. to their root *go*.
 - Stemmers are freely available for many languages.
- Latent semantic indexing.
 - Similar to principal component analysis.
 - Map instances into new coordinate system.
 - New coordinates correspond to semantic concepts.
 - Reduce dimensionality by neglecting coordinates with low variance (= hardly occurring semantic concepts).

Latent Semantic Indexing

	d_1	d_2	d_3	d_4	d_5	d_6
cosmonaut	1	0	1	0	0	0
astronaut	0	1	0	0	0	0
moon	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

- Term-document matrix A .
- Find matrices T , S , and D such that $A = T \times S \times D^T$.
 - ① Compute eigenvalues e_1, \dots, e_p of $A^T A$.
 - ② Compute matrix D comprising the corresponding eigenvectors.
 - ③ Define $S = \text{diag}(e_1, \dots, e_p)$.
 - ④ Compute T , for instance by Gram-Schmidt orthogonalization.

Applications

Classification of Text Documents

- Annotate text documents with class labels.
 - Binary classification.
 - Multi-class classification.
 - Multi-label classification.
- Applications:
 - Detect spam messages (binary).
 - Classify web pages into web directories (multi-class).
 - Classify news articles (multi-label).
- Learn a classifier from labeled documents.
 - For text linear classifiers have been proven to perform well.
 - E.g., linear support vector machines.

Support Vector Machines

- Binary text classification (e.g., ham vs. spam).
- SVMs minimize upper bound on regularized empirical risk.
- Labeled documents $\{(d_i, y_i)\}_{i=1}^m$ with $y_i \in \{+1, -1\}$.

$$\min_{w, b, \xi} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i$$

$$\text{s.t.} \quad \forall_{i=1}^m : \quad y_i(\langle w, \phi(d_i) \rangle + b) \geq 1 - \xi_i$$

$$\forall_{i=1}^m : \quad \xi_i \geq 0.$$

- Document representation $\phi(d)$.
- Easily generalized to multi-class and multi-label problems.
 - Strategy: one-against-one, one-against-all.

Evaluation of Text Classifiers

- Misclassification error rates not appropriate when $P(+1)$ small.
 - E.g., how good is an error of 5% when $P(+1) = 3\%$?
- Solution: measure performance of decision function $f(x)$
- Precision/Recall

$$\text{Precision}(f) = P(y = +1 | f(x) = +1) = \frac{TP}{TP + FP}$$

$$\text{Recall}(f) = P(f(x) = +1 | y = +1) = \frac{TP}{TP + FN}$$

- Breakeven point: $\text{Prec}(f) = \text{Rec}(f)$, F -measure: $F = \frac{2 \cdot \text{Prec}(f) \cdot \text{Rec}(f)}{\text{Prec}(f) + \text{Rec}(f)}$.
- Receiver Operating Characteristic (ROC).
 - Area under the ROC curve: $\text{AUC}(f) = P(f(x_{\text{pos}}) > f(x_{\text{neg}}))$.

Experiment

- Reuters-21578 data set (ModApte compilation).
 - News articles from Reuters news archive.
- 9603 training documents, 3299 test documents, 90 classes.
- Preprocessing: 9962 distinct terms in dictionary.
- Features: normalized term frequencies.
- Baselines: naïve Bayes, C4.5, Rocchio, k -nearest neighbors.

E.D. And F. MAN TO BUY INTO HONG KONG FIRM

The U.K. Based commodity house E.D. And F. Man Ltd and Singapore's Yeo Hiap Seng Ltd jointly announced that Man will buy a substantial stake in Yeo's 71.1 pct held unit, Yeo Hiap Seng Enterprises Ltd. Man will develop the locally listed soft drinks manufacturer into a securities and commodities brokerage arm and will re-name the firm Man Pacific (Holdings) Ltd.

Empirical Results

	Bayes	Rocchio	C4.5	k-NN	SVM (poly) degree $d =$					SVM (rbf) width $\gamma =$			
					1	2	3	4	5	0.6	0.8	1.0	1.2
earn	95.9	96.1	96.1	97.3	98.2	98.4	98.5	98.4	98.3	98.5	98.5	98.4	98.3
acq	91.5	92.1	85.3	92.0	92.6	94.6	95.2	95.2	95.3	95.0	95.3	95.3	95.4
money-fx	62.9	67.6	69.4	78.2	66.9	72.5	75.4	74.9	76.2	74.0	75.4	76.3	75.9
grain	72.5	79.5	89.1	82.2	91.3	93.1	92.4	91.3	89.9	93.1	91.9	91.9	90.6
crude	81.0	81.5	75.5	85.7	86.0	87.3	88.6	88.9	87.8	88.9	89.0	88.9	88.2
trade	50.0	77.4	59.2	77.4	69.2	75.5	76.6	77.3	77.1	76.9	78.0	77.8	76.8
interest	58.0	72.5	49.1	74.0	69.8	63.3	67.9	73.1	76.2	74.4	75.0	76.2	76.1
ship	78.7	83.1	80.9	79.2	82.0	85.4	86.0	86.5	86.0	85.4	86.5	87.6	87.1
wheat	60.6	79.4	85.5	76.6	83.1	84.5	85.2	85.9	83.8	85.2	85.9	85.9	85.9
corn	47.3	62.2	87.7	77.9	86.0	86.5	85.3	85.7	83.9	85.1	85.7	85.7	84.5
microavg.	72.0	79.9	79.4	82.3	84.2	85.1	85.9	86.2	85.9	86.4	86.5	86.3	86.2
					combined: 86.0					combined: 86.4			

Fig. 2. Precision/recall-breakeven point on the ten most frequent Reuters categories and microaveraged performance over all Reuters categories. k -NN, Rocchio, and C4.5 achieve highest performance at 1000 features (with $k = 30$ for k -NN and $\beta = 1.0$ for Rocchio). Naive Bayes performs best using all features.

- SVMs well suited for sparse, high-dimensional feature spaces. (Joachims, 1998)

Summary & Further Applications

Summary

- Characteristics of natural language.
 - Infinitely many terms.
 - Ambiguous.
 - Disambiguation by context information.
- Document representations:
 - Bag-of-words, TF.IDF, n -grams.
 - Relevant for classification, clustering, and ranking tasks.
 - Dimensionality reduction techniques.
- Exemplary application.
 - Text classification with SVMs.
 - Performance measures.

Further Applications

- Potentially more challenging high-level tasks:
 - Natural language parsing.
 - Named entity recognition.
 - Named entity resolution.
 - Machine translation.
 - Sentiment prediction.
 - Document summarization.
 - Question answering.
 - ...

Literatur/Referenzen

- R. Baeza-Yates & B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.
- X. Huang, A. Acero & H.-W. Hon, *Spoken Language Processing*, Prentice Hall, 2001.
- T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, *Proceedings of the European Conference on Machine Learning (ECML)*, 1998.
- T. Joachims, A statistical learning model of text classification for support vector machines, *Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR)*, 2001.
- T. Joachims, *Learning to Classify Text with Support Vector Machines*, Kluwer Academic Publishers / Springer, 2002.
- C. D. Manning & H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 2003.