# Appendix A: Religion Data

## A.1 The Sub-corpora

| | Buddhism | Christianity | Hinduism | Islam | Judaism |
|---|---|---|---|---|---|
| Original corpus size | 1.44 (8.66) | 1.89 (10.77) | 2.14 (12.99) | 1.51 (8.72) | 1.56 (9.18) |
| Lemmatized corpus | 0.76 (5.57) | 0.98 (6.71) | 1.20 (8.63) | 0.77 (5.38) | 0.83 (5.87) |
| size in millions of word tokens (megabytes) | | | | | |
| # documents | 58 | 44 | 44 | 52 | 44 |
| Encyclopedic entries | 8 | 8 | 8 | 8 | 8 |
| FAQ files | 4 | 4 | 4 | 2 | 5 |
| Online periodicals | 4 | 4 | 1 | 6 | 5 |
| Other web-sites / online books | 42 | 28 | 31 | 36 | 26 |

- The sub-corpora used for this work contain the sub-corpora used by Marx, Dagan, Buhmann and Shamir (2002), extending them by 25-50%.
- Inclusive documents have been chosen: ones that essentially refer to a religion as a whole, rather than being pre-limited to a specific aspect or school. This way, we have tried to create sub-corpora of general character that repeatedly refer to a variety of aspects on roughly the same level of detail. This makes the practice of filtering out key-terms appearing on less than four documents somewhat more reliable.
- Some of the documents have been created as a merge of several pages appearing on the same Internet site. For example, each "online-periodical" document consists of up to tenth of individual articles.
- We have made efforts to exclude texts that literally repeat over different web pages.
- The use of a POS tagger and lemmatizer is also new relatively to Marx et al. (2002). There, filtering of function words was controlled by a pre-determined list. Here, we have identified content words by their part of speech, leaving only the lemmatized nouns (including names), verbs, adjectives and adverbs. Numbers tagged as cardinal or ordinal numbers were replaced by *~card~* and *~ord~* signs
- Some of the most prevalent alternative part-of-speech tags have been attached to the lemmatized word tokens. For example, *mean/V* stands for occurrences of the lemma 'mean' tagged as a verb, while *mean* stands for the noun sense, which is more prevalent within our sub-corpora, as well as for any other part of speech that the tagger has (erroneously) attached to the same lemma. Other tags are */N* for names or nouns and */J* for adjectives and adverbs. The alternative tags are attached whenever the alternative repeats 50 times or more (in all sub-corpora).

## A.2 The Features

For each specific cross religion comparison, the features used – i.e., counted content words co-located with the clustered key-terms – are those occurring in both compared sub-corpora (at least twice in each corpus). The numbers of features used for each comparison:

| | | | |
|---|---|---|---|
| Common to all 5 religions: | 6796 | Christianity, Islam and Judaism: | 7768 |
| Buddhism and Christianity: | 8717 | Buddhism and Hinduism: | 9905 |
| Buddhism and Islam: | 7973 | Buddhism and Judaism: | 8735 |
| Christianity and Hinduism: | 10410 | Christianity and Islam: | 8604 |
| Christianity and Judaism: | 9641 | Hinduism and Islam: | 9438 |
| Hinduism and Judaism: | 10454 | Islam and Judaism: | 8563 |

Following is the list of the 58 features that are among the 100 most common features in at least four of the five sub-corpora. The numbers in brackets indicate the number of joint occurrences in which the feature is involved (in order to calculate $p(y)$ for any feature $y$ within the configuration incorporating all sub-corpora, one should divide the number in brackets by the total count of co-occurrences in all sub-corpora):

**have** (99303), **not** (71697), **god** (54503), **do** (42993), **one** (36608), **say** (30779), **life** (25722), **man** (20262), **other** (20243), **people** (20219), **make** (19733), **only** (19065), **give** (18919), **come** (17978), **world** (17579), **so** (17402), **time** (17059), **also** (17020), **being** (16685), **~card~** (15591), **many** (15433), **as** (15258), **way** (15077), **know** (15029), **more** (14830), **see** (13531), **then** (13507), **word** (13203), **day** (13160), **go** (13115), **first** (13014), **take** (12690), **most** (12570), **become** (12336), **good** (12277), **even** (12273), **great** (12037), **believe** (11307), **human** (10786), **call** (10656), **out** (10527), **year** (10076), **live** (9990), **find** (9937), **own** (9877), **such** (9575), **use/V** (9260), **book** (8928), **very** (8889), **up** (8880), **two** (8875), **person** (8411), **mean/V** (8386), **now** (8271), **state** (8194), **same** (8149), **bring** (8135), **teach** (8050).

The total co-occurrence count in the corpus (all sub-corpora): **4892150**

The rest of the 100 most common features within each individual corpus follow. The numbers in brackets indicate the co-occurrence count within the individual corpus. In order to calculate feature probability conditioned on the corpus $p(y|X_i)$ for any feature $y$ one would divide the bracketed number by the total number of co-occurrences within the corpus, which appear at the end of each list.

Buddhism

**buddhist** (8452), **buddha** (7601), **buddhism** (6470), **practice** (3344), **teaching** (3236), **mind** (3036), **meditation** (2307), **monk** (2068), **path** (1949), **suffering** (1869), **tradition** (1866), **thing** (1774), **right** (1733), **truth** (1717), **develop** (1678), **just** (1622), **religious** (1620), **death** (1586), **action** (1569), **understand** (1538), **religion** (1490), **enlightenment** (1468), **sense** (1467), **follow** (1448), **teacher** (1439), **lead** (1430), **nature** (1426), **order** (1421), **think** (1385), **three** (1372), **other/N** (1357), **term** (1352), **text** (1343), **spiritual** (1295), **experience** (1284), **form** (1269), **different** (1214), **part** (1202), **arise** (1200), **existence** (1199), **sangha** (1174), **well** (1168), **school** (1165), **four** (1163).

The total co-occurrence count in the corpus: **812850**

Christianity

**jesus** (15446), **christ** (14275), **church** (7144), **lord** (6303), **sin** (5438), **son** (5041), **holy** (4918), **bible** (4260), **faith** (4226), **thing** (3833), **christian** (3740), **father** (3693), **christian/J** (3251), **save** (3131), **gospel** (3127), **death** (2993), **heaven** (2954), **tell** (2803), **speak** (2803), **power** (2700), **work** (2655), **just** (2609), **paul** (2539), **heart** (2513), **prayer** (2472), **john** (2430), **salvation** (2415), **write** (2393), **name** (2339), **get** (2334), **baptism** (2277), **think** (2268), **ever** (2240), **body** (2232), **receive** (2211), **love** (2151), **law** (2144), **grace** (2144), **child** (2143), **therefore** (2078), **holy/J** (2067), **scripture** (2062), **testament** (2035), **want** (2021), **die** (2009).

The total co-occurrence count in the corpus: **1281079**

Hinduism

**hindu** (10614), **india** (5554), **hinduism** (4262), **religion** (3975), **temple** (3529), **religious** (2854), **spiritual** (2835), **indian** (2748), **yoga** (2672), **worship** (2612), **lord** (2535), **child** (2387), **soul** (2204), **ancient** (2160), **family** (2097), **culture** (2067), **vedic** (1943), **sri** (1864), **body** (1835), **mind** (1774), **include** (1726), **part** (1725), **form** (1689), **school** (1681), **system** (1644), **swami** (1627), **krishna** (1625), **philosophy** (1617), **tradition** (1607), **knowledge** (1590), **different** (1586), **dharma** (1581), **nature** (1534), **vedas** (1519), **karma** (1502), **practice** (1480), **sacred** (1475), **new** (1473), **place** (1452), **ritual** (1449), **today** (1447), **scripture** (1423), **high** (1420).

The total co-occurrence count in the corpus: **1002100**

Islam

**allah** (12391), **prophet** (8795), **islam** (7799), **muhammad** (4130), **muslims** (3980), **messenger** (3254), **islamic** (3097), **religion** (2994), **follow** (2663), **muslim** (2483), **law** (2207), **woman** (2132), **belief** (2079), **worship** (2078), **faith** (2058), **reveal** (1975), **prayer** (1973), **muslim/N** (1965), **ask** (1903), **knowledge** (1844), **peace** (1843), **verse** (1842), **heart** (1749), **holy** (1685), **true** (1658), **revelation** (1612), **jesus** (1587), **order** (1573), **accord** (1570), **create** (1484), **name** (1442), **qur~an** (1429), **fact** (1406), **accept** (1367), **fast** (1359), **send** (1357), **thing** (1333), **lord** (1333), **message** (1322), **truth** (1319), **right/N** (1300), **place** (1293), **last** (1292), **believer** (1287), **well** (1275).

The total co-occurrence count in the corpus: **913241**

Judaism

**jewish** (10726), **torah** (6557), **rabbi** (4150), **judaism** (3666), **jews** (3541), **law** (3482), **israel** (3456), **woman** (2562), **jew** (2328), **moses** (1940), **child** (1935), **name** (1845), **prayer** (1793), **community** (1738), **temple** (1723), **religious** (1706), **write** (1687), **commandment** (1604), **create** (1585), **part** (1569), **tell** (1514), **land** (1511), **begin** (1507), **just** (1438), **include** (1421), **place** (1396), **talmud** (1352), **spiritual** (1351), **speak** (1350), **synagogue** (1348), **tradition** (1336), **new** (1330), **father** (1313), **reform** (1305), **movement** (1275), **service** (1259), **accord** (1238), **abraham** (1238), **well** (1206), **rabbinical** (1206), **understand** (1199), **jerusalem** (1192).

The total co-occurrence count in the corpus: **881569**

## A.3 The Clustered Keyword Sets

The sizes of the keyword sets are as follows:

Buddhism – 227; Christianity – 235; Hinduism – 177; Islam – 221; Judaism – 232.

The whole sets can be seen in the results Appendix E. Here, the clustered keyword sets are exemplified through arbitrarily chosen ~10% subsets (we have picked elements number 1, 11, 21, … and so on, according to their lexicographic order). Along with each element, we indicate its three most prominent features that are not included in any of the above lists of common 100 features per corpus. The individual feature lists below demonstrate well the information conveyed by the thousands of features that are not very frequent. Each feature is preceded by its relative rank in terms of number of co-occurrences with the particular element. The bracketed numbers, that is the count of joint occurrences, divided by the total co-occurrence count of the element $x$, gives the respective conditional probabilities $p(y|x)$. The total number of $x$ co-occurrences divided by the total number of co-occurrences within the corpus (or within all sub-corpora), which has been indicated previously, gives the probability $p(x|X_i)$ (or $p(x)$).

Buddhism

| Key-term | Features (co-occurrence count) | | | Total count |
|---|---|---|---|---|
| **Abbot** | 2.monastery (13) | 4.here (7) | 7.there (5) | 461 |
| **Asceticism** | 4.five (33) | 8.wander (20) | 13.austerity (16) | 2307 |
| **Being** | 6.sentient (200) | 13.living (122) | 35.happiness (77) | 24953 |
| **Burma** | 1.thailand (36) | 2.lanka (35) | 5.cambodia (15) | 795 |
| **conditioning** | 3.process (9) | 4.consciousness (8) | 5.condition (7) | 544 |
| **Discipline** | 2.rule (31) | 6.monastic (27) | 7.code (21) | 2683 |
| **Emptiness** | 3.phenomenon (25) | 5.realize (22) | 8.realization (20) | 2132 |
| **Family** | 6.friend (30) | 10.leave (24) | 11.member (20) | 3239 |
| **full~moon** | 1.month (26) | 3.moon (15) | 5.night (12) | 505 |
| **Hinduism** | 15.caste (13) | 23.principle (9) | 24.orthodox/J (9) | 2194 |
| **Karma** | 5.result (91) | 6.bad (79) | 8.rebirth (71) | 7730 |
| **Liberation** | 8.achieve (20) | 9.insight (20) | 12.attain (19) | 2754 |
| **Meet** | 10.group (17) | 13.need (16) | 28.attend (12) | 3041 |
| **noble~truths** | 6.suffer (47) | 8.noble (44) | 10.cause (36) | 2937 |
| **Phenomenon** | 1.mental (34) | 3.emptiness (29) | 4.physical (27) | 2472 |
| **psychologist** | 2.modern (7) | 4.philosopher (4) | 5.view (3) | 311 |
| **Robe** | 1.wear (35) | 4.bowl (18) | 5.yellow (14) | 1297 |
| **Sanskrit** | 1.pali (44) | 5.language (24) | 9.translate (17) | 1616 |
| **Society** | 5.individual (32) | 7.social (28) | 14.member (20) | 4698 |
| **Student** | 4.western (18) | 10.master (13) | 14.zen (11) | 1806 |
| **Text** | 4.pali (67) | 6.early (55) | 14.group (36) | 7680 |
| **Universe** | 8.phenomenon (19) | 16.everything (16) | 18.entire (15) | 2566 |
| **Wisdom** | 2.compassion (100) | 5.perfection (63) | 12.virtue (38) | 6313 |

Christianity

| Key-term | Features (co-occurrence count) | | | Total count |
|---|---|---|---|---|
| **Abraham** | 2.promise (91) | 4.seed (50) | 10.isaac (32) | 2866 |
| **Association** | 4.unitarian (9) | 7.evangelical (9) | 8.universalist (8) | 668 |
| **Believing** | 32.baptize (75) | 43.ye (60) | 50.reason (55) | 19264 |
| **Cardinal** | 2.bishop (13) | 3.pope (10) | 4.elect (9) | 285 |
| **Confess** | 11.forgive (30) | 19.mouth (22) | 27.faithful (15) | 3468 |
| **Doctrinal** | 8.trinity (67) | 21.christianity (39) | 26.principle (34) | 8561 |
| **Evangelical** | 2.theology (67) | 19.century (22) | 22.group (19) | 5851 |
| **Fire** | 1.lake (68) | 4.hell (48) | 6.burn (39) | 3234 |
| **Gift** | 18.tongue (37) | 20.christmas (35) | 21.prophecy (33) | 5455 |
| **Heaven** | 4.earth (268) | 14.kingdom (100) | 18.hell (84) | 12142 |
| **Idolatry** | 4.forme (4) | 12.note (3) | 14.inordinate (3) | 335 |
| **Jew** | 8.gentile (40) | 12.christianity (31) | 17.roman (23) | 5074 |
| **Law** | 9.keep (72) | 24.divine/J (39) | 25.break (38) | 9470 |
| **Mary** | 3.virgin/N (40) | 6.mother (33) | 7.joseph (28) | 2248 |
| **Moral** | 5.goal (38) | 8.evil/N (33) | 10.acceptable (25) | 4218 |
| **Passage** | 11.refer (29) | 12.read (25) | 14.meaning (22) | 3843 |
| **Pope** | 5.ii (39) | 6.bishop (33) | 7.roman (25) | 2273 |
| **Question** | 6.answer/V (118) | 7.answer (106) | 10.raise (45) | 8434 |
| **Revelation** | 16.chapter (19) | 22.special (16) | 37.divine/J (12) | 3480 |
| **Salvation** | 21.plan (52) | 28.eternal (43) | 32.necessary (37) | 9720 |
| **Soul** | 8.win (64) | 20.winner (45) | 25.immortal (36) | 7894 |
| **Teach** | 11.pray (56) | 12.doctrine (56) | 31.disciple (26) | 7404 |
| **Trinity** | 2.doctrine (66) | 16.incarnation (17) | 20.unity (12) | 2160 |
| **Worship** | 10.music (47) | 24.sunday (30) | 27.praise (28) | 7227 |

Hinduism

| Key-term | Features (co-occurrence count) | | | Total count |
|---|---|---|---|---|
| **Advaitha** | 8.theory (5) | 13.pure (4) | 15.doctrine (4) | 419 |
| **Attain** | 4.liberation (51) | 12.eternal (32) | 13.bliss (32) | 4421 |
| **brahma~sutra** | 1.commentary (21) | 2.upanishads (15) | 3.gita (11) | 293 |
| **classical** | 1.music (85) | 3.dance (58) | 9.sanskrit (16) | 1942 |
| **divinity** | 9.mother (81) | 17.self (59) | 22.society (48) | 12320 |
| **festival** | 4.celebrate (51) | 7.annual (38) | 9.hold (32) | 3622 |
| **gita** | 2.upanishads (49) | 10.commentary (23) | 17.sutra (20) | 3349 |
| **hymn** | 2.veda (41) | 3.collection (30) | 4.rig (29) | 2263 |
| **karma** | 10.reincarnation (78) | 11.bad (73) | 13.past (62) | 9117 |
| **mahabharata** | 1.epic (73) | 3.gita (23) | 7.puranas (18) | 1788 |
| **philosophy** | 23.six (42) | 35.science (34) | 37.upanishads (33) | 10430 |
| **question** | 3.answer (83) | 4.answer/V (72) | 13.scend (25) | 4432 |
| **rigveda** | 2.hymn (46) | 3.veda (35) | 8.old (19) | 1929 |
| **scholar** | 3.sanskrit (30) | 4.western (30) | 14.leader (15) | 2707 |
| **smriti** | 4.manu (15) | 5.remember (13) | 6.literature (12) | 784 |
| **student** | 5.university (55) | 15.learn (29) | 20.college (25) | 5771 |
| **tradition** | 22.value (52) | 33.art (44) | 44.preserve (34) | 14172 |
| **west** | 5.east (93) | 12.astrology (63) | 21.eastern (42) | 11606 |

Islam

| Key-term | Features (co-occurrence count) | | | Total count |
|---|---|---|---|---|
| **abraham** | 7.noah (54) | 10.ishmael (39) | 12.adam (36) | 3195 |
| **army** | 3.battle (16) | 6.fight (10) | 10.commander (9) | 1648 |
| **bible** | 20.mention (7) | 25.statement (7) | 29.difference (6) | 1306 |
| **christian** | 35.doctrine (24) | 44.trinity (21) | 46.missionary (21) | 7241 |
| **creature** | 7.creator (28) | 10.mercy (22) | 11.universe (18) | 1880 |
| **earth** | 16.creation (34) | 20.face (26) | 21.belong (26) | 5823 |
| **faith** | 20.article (52) | 26.deed (48) | 44.reject (28) | 9806 |
| **food** | 1.eat (57) | 4.drink (33) | 10.abstain (17) | 2239 |
| **guide** | 12.mankind (54) | 22.clear (36) | 24.seek (35) | 6521 |
| **house** | 5.enter (32) | 7.build (28) | 11.pilgrimage (24) | 2838 |
| **ishmael** | 4.isaac (20) | 7.jacob (17) | 8.build (15) | 854 |
| **kaabah** | 1.mecca (23) | 4.build (17) | 6.house (16) | 996 |
| **marriage** | 6.wife (58) | 8.divorce (42) | 16.husband (26) | 3801 |
| **mother** | 8.wife (30) | 12.sister (26) | 23.baby (15) | 2758 |
| **pilgrimage** | 1.hajj (99) | 2.mecca (77) | 5.duty (38) | 3113 |
| **purification** | 5.wealth (18) | 23.alms (7) | 24.intention (7) | 1343 |
| **responsibility** | 11.hold (22) | 12.social (21) | 14.society (20) | 3145 |
| **science** | 6.modern (34) | 7.technology (33) | 19.art (18) | 4308 |
| **social** | 2.economic (78) | 4.political (72) | 8.society (49) | 4999 |
| **submission** | 5.obedience (44) | 9.total (24) | 13.complete (17) | 2005 |
| **testimony** | 6.confirm (12) | 7.bear (12) | 12.ramadan (8) | 704 |
| **water** | 7.drink/V (21) | 8.jug (19) | 12.down (16) | 2489 |
| **writing** | 12.read (5) | 15.jail (4) | 16.leaf (4) | 685 |

Judaism

| Key-term | Features (co-occurrence count) | | | Total count |
|---|---|---|---|---|
| **abraham** | 2.isaac (92) | 5.sarah (66) | 6.jacob (52) | 5074 |
| **ashkenazim** | 8.jewry (10) | 12.germany (9) | 15.custom (7) | 1094 |
| **canaan** | 2.egypt (18) | 7.israelite (7) | 9.jacob (7) | 674 |
| **command** | 12.keep (17) | 14.love/V (16) | 17.sanctify (16) | 2980 |
| **discuss** | 4.issue (41) | 14.debate (16) | 15.chapter (16) | 3951 |
| **europe** | 2.eastern/N (42) | 5.western/N (17) | 6.century (16) | 1476 |
| **family** | 4.member (71) | 5.friend (62) | 14.home (31) | 6527 |
| **gemara** | 2.commentary (14) | 3.together (8) | 7.answer (7) | 620 |
| **hebrew** | 6.language (40) | 11.union (29) | 12.hebraic (28) | 4374 |
| **humanity** | 11.creation (7) | 14.history (6) | 20.image (5) | 1046 |
| **jesus** | 4.messiah (18) | 15.consider (7) | 19.individual (6) | 1001 |
| **law** | 7.oral (126) | 14.custom (75) | 24.code (63) | 17337 |
| **member** | 8.congregation (32) | 12.committee (21) | 13.group (18) | 3391 |
| **mourn** | 2.period (36) | 8.house (13) | 12.destruction (9) | 1088 |
| **people** | 27.choose (87) | 38.egypt (73) | 47.covenant (62) | 26564 |
| **rabbi** | 8.ben (99) | 10.congregation (98) | 14.role (89) | 21829 |
| **revelation** | 2.sinai (27) | 6.creation (14) | 8.divine/J (13) | 1658 |
| **salvation** | 5.miracle (8) | 9.redemption (6) | 12.covenant (6) | 776 |
| **service** | 7.healing (54) | 9.morning (50) | 16.attend (35) | 6049 |
| **spirit** | 4.healing (35) | 11.evil (16) | 26.letter (9) | 2535 |
| **talit** | 1.wear (34) | 4.shawl (12) | 5.corner (11) | 714 |
| **text** | 6.biblical (46) | 7.read (40) | 9.meaning (31) | 5107 |
| **wisdom** | 10.understanding (19) | 15.solomon (15) | 26.divine (10) | 2555 |
| **zionism** | 10.secular (13) | 12.political (12) | 16.century (10) | 1478 |