**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI**

**INSTITUT TEKNOLOGI SEPULUH NOPEMBER**

**USULAN TUGAS AKHIR**

# IDENTITAS PENGUSUL

**NAMA : Syahrul Munif**

**NRP : 5110 100 210**

**DOSEN WALI : Dr.techn. Ir.Raden Venantius Hari Ginardi, M.Sc.**

**DOSEN PEMBIMBING : 1.Dr.Chastine Fatichah, S.Kom, M.Kom.   
 2.**

# JUDUL TUGAS AKHIR

“Penerapan Metode *Heuristic Semi-supervised Fuzzy Co-clustering Algorithm with Ruspini’s Condition* (SS-HFCR) Untuk Pengelompokan Dokumen Teks”

# LATAR BELAKANG

Dokumen merupakan sebuah tulisan yang memuat informasi [[1](#Wik13)]. Pada zaman dahulu,dokumen identik dengan informasi yang ditulis di kertas,bambu,kulit hewan dll. Pada zaman sekarang,dokumen tidak hanya yang berbentuk nyata atau *hardcopy* seperti buku,surat dan lain-lain. Dokumen pada zaman sekarang lebih banyak berbentuk digital/*softcopy*. Media penyimpanan digital untuk dokumen ini juga semakin besar. Sehingga semakin banyak dokumen yang tersimpan. Pada suatu institusi atau lembaga, banyaknya dokumen bisa menjadi suatu masalah tersendiri. Dikarenakan banyaknya dokumen,menyusun dokumen supaya terkelompok sesuai topik menjadi sulit. Memilah secara manual dapat dilakukan untuk dokumen-dokumen ini. Cara ini mungkin bisa dilakukan jika jumlah dokumen sedikit. Jika dokumen berjumlah terlalu banyak,cara ini tidak bisa di lakukan.

*Machine learning* adalah sekumpulan metode yang dapat digunakan untuk mengenali pola dari skumpulan data,serta mampu memprediksi data yang akan keluar pada masa datang atau digunakan untuk mendukung keputusan dibawah ketidakpastian [[2](#Wik131)]. *Document clustering* merupakan bagian dari *machine learning* yang dapat digunakan untuk mengatur dokumen supaya bisa terkelompok dalam bagian-bagian grup [[3](#Yan13)].

Jumlah data yang besar sering menjadikan permasalahan tersendiri dalam pengelompokan. Selain itu,adanya data yang rusak dan data yang saling tumpang tindih [[3](#Yan13)]. Beberapa algoritma pengelompokan di kembangkan untuk menyelesaikan permasalahan pengelompokan ini. Contoh beberapa algoritma yang dikembangkan adalah *co-clustering,fuzzy clustering,fuzzy co-clustering,*pengelompokan berdasar konsep matriks (NMF) dan lain sebagainya. Setiap metode mempunyai ciri tersendiri,misalnya *co-clustering* umumnya baik untuk menangani data yang besar dengan cara secara serentak mengelompokkan dokumen dan kata berdasarkan tingginya hubungan antara keduanya. *Fuzzy clustering* digunakan untuk permasalahan data yang saling tumpang tindih,tetapi menjadi tidak bagus ketika terdapat anomali data. Di dalam *fuzzy co-clustering* keterkaitan direpresentasikan dengan derajat keanggotaan dokumen terhadap setiap kelompok/kluster. *Fuzzy co-clustering* merupakan variasi terbaru dari *fuzzy c-means* dan pendekatan *co-clustering* yang bekerja pada dataset tertentu*.*

Masih menjadi tantangan tersendiri untuk mengembangkan metode pengelompokan terhadap dataset yang lebih kompleks dan murni metode tanpa pengawasan pengguna (*unsupervised)* [[3](#Yan13)]*.* Metode yang masih menggunakan pengetahuan terdahulu yang melibatkan pengguna dan metode tanpa pengawasan disebut metode *semi-supervised.* Pengetahuan terdahulu yang tersedia meningkatkan akurasi metode. Metode *semi-supervised* yang ada sekarang ini pada umumnya terbagi menjadi dua yaitu berdasar adaptasi kemiripan dan metode berdasar pencarian.

Dalam metode adaptasi kemiripan nilai kemiripan beberapa objek diberikan. Pengguna umumnya melakukan pelatihan terhadap data yang ada dengan menggunakan pembelajaran jarak matriks. Pada metode berdasar pencarian,algoritma ini dirancang untuk menggunakan label yang sudah ada atau berbagai aturan yang mengarahkan dalam pengelompokan yang sesuai. Menyediakan beberapa data yang sudah terkelompokan merupakan pendekatan yang mudah sebagai inisiasi kelompok / *initial cluster.* Namun,proses pemberian label menjadi sulit untuk setiap kelompok dalam kasus nyata,seperti pengelompokan data GPS. Kebanyakan metode *semi-supervised* lebih terfokus dalam membuat aturan berpasangan *must link /* harus terhubung dan *can not link* / tidak dapat terhubung.

*Heuristic Semi-supervised Fuzzy Co-clustering Algorithm with Ruspini’s Condition* (SS-HFCR) merupakan salah satu teknik baru dalam *document clustering.* Metode ini menggabungkan metode *fuzzy clustering, co-clustering* dan pengelompokan *semi-supervised.* SS-HFCR menggunakan pengetahuan yang sebelumnya ada dalam bentuk batasan/aturan berpasangan untuk domain dokumen. Setiap aturan secara spesifik mengatur apakah pasangan dokumen mempunyai aturan “harus/*must*” atau “tidak dapat/*cannot*” dikelompokan dalam satu kelompok. Tujuan dari algotima ini adalah untuk meningkatkan akurasi pengelompokan,menekan kepekaan terhadap parameter *fuzzi* dengan pengetahuan terdahulu yang terbatas sedangkan kompleksitas algoritma tetap rendah. SS-HFCR digunakan dalam menentukan suatu dokumen termasuk dalam suatu topik. Diharapkan dengan metode/teknik ini pengelompokan dokumen menjadi lebih mudah.

# RUMUSAN MASALAH

Rumusan masalah yang diangkat dalam tugas akhir ini adalah:

1. Bagaimana mengimplementasikan metode SS-HFCR?
2. Bagaimana metode SS-HFCR dapat digunakan untuk mengelompokkan dokumen teks?
3. Bagaimana mengevaluasi kinerja metode SS-HFCR ?

# BATASAN MASALAH

Permasalahan yang dibahas dalam tugas akhir ini memiliki beberapa batasan, yaitu sebagai berikut :

1. *Dataset* yang digunakan adalah *20 Newsgroups,* *Reuters-21578 R8,* dan *Webkb* *Reuters-21578 [[1]](#footnote-1).*
2. Metode yang digunakan adalah *Heuristic Semi-supervised Fuzzy Co-clustering Algorithm* (SS-HFCR).
3. Evaluasi yang dilakukan melingkupi akurasi,kestabilan dan efesiensi.

# TUJUAN PEMBUATAN TUGAS AKHIR

Tugas akhir ini bertujuan untuk :

1. Mengimplementasikan metode SS-HFCR .
2. Mengelompokkan dokumen teks dengan metode SS-HFCR.
3. Mengevaluasi kinerja metode SS-HFCR berdasarkan akurasi,kestabilan,dan efesiensi.

# MANFAAT TUGAS AKHIR

Manfaat yang diharapkan dari tugas akhir ini adalah Mengaplikasikan kakas bantu untuk mempermudah dalam mengelompokkan/pengelompokan dokumen teks.

# TINJAUAN PUSTAKA

**8. 1 *Heuristic Semi-supervised Fuzzy Co-clustering Algorithm with Ruspini’s Condition* (SS-HFCR)**

Algoritma *semi-supervised heuristik semi-supervised fuzzy co-clustering with ruspini’s condition* (SS-HFCR) digunakan untuk mengelompokkan data dokumen web yang besar. Di dalam pendekatan ini,proses pengelompokan di lakukan dengan memasukan beberapa pengetahuan sebelumnya dalam bentuk aturan berpasangan yang disediakan oleh pengguna ke kerangka *fuzzy co-clustering* [[3](#Yan13)]. Setiap aturan menentukan apakah pasangan dokumen "harus terhubung" atau "tidak dapat terhubung" di masukan dalam klaster yang sama. Selanjutnya,diformulasikan fungsi biaya aglomerasi kompetitif yang juga dapat digunakan untuk pengetahuan sebelumnya di dalam proses pengelompokan. Pembelajaran berdasar eksperimen dengan *dataset* berjumlah besar menunjukkan kekuatan dan potensi dari SS-HFCR dalam sisi akurasi,kestabilan,dan efesiensi.

|  |  |
| --- | --- |
| **Notasi** | **Arti** |
| **D** | Asosiasi matriks kata dalam dokumen |
| *c/C* | Indeks kelompok/Jumlah kelompok |
| M | Jumlah kata |
| N | Jumlah dokumen |
| *xi* | Dokumen ke i dalah *R*­M |
| *dij* | *tf-idf* dari kata j dalam dokumen i |
| **U,V** | Dokumen dan kata matrik *fuzzy* |
| *uci,vcj* | Nila derajat keanggotaan dokumen dan kata |
| Tu*,*Tv | Derajat keanggotaan *fuzzy* yang ditentukan pengguna |
| T*d* | Bobot faktor(harga penalti) dari *constraint* |

Fungsi objektif dari SS-HFCR dapat dituliskan sebagai berikut:

Dengan optimasi fungsi *J* sesuai dua aturan sebagai berikut:

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . (1)

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . (2)

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . (3)

Ini dinamakan pendekatan *fuzzy co-clustering,*setiap dokumen *i* di berikan derajat kecenderungan keanggotaan *fuzzy uci* terhadapkelompok *c.* Untuk setiap kata *j* dalam *dataset* diberikan derajat kecenderungan fuzzy *vcj* terhadap kelompok c. Keduanya harus digunakan dalam fungsi objektif. Korelasi antara kata dan dokumen haruslah kuat dalam satu grup/kelompok. Sehingga, bagian pertama yang disebut derajat agregasi haruslah semaksimal mungkin terhadap suatu klaster. Sehingga membuat keterkaitan yang tinggi antara dokumen dan kata ( diindikasikan *dij* tinggi) menjadi satu klaster. Hal itu di lakukan atas dasar pemikiran bahwa *co-cluster* yang baik seharusnya mempunyai koherensi ikatan yang kuat antar anggotanya.

Memaksimalkan derajat agregasi mungkin akan mendapatkan hasil pengelompokkan yang tidak diinginkan. Misalkan bagian pertama persamaan adalah **G:D =** ,*gij*(Setiap elemen **G**) dinyatakan . Nilai dapat bervariasi antara 0 dan *NK*. Variasi ini berimplikasi,memaksimalkan derajat agregasi dalam kasus ini akan mengaburkan pembentukan *co-cluster* dengan besarnya nilai . Nilai tadi tidak bergantung pada jumlah pastisi **D.** Maka dari itu,*co-cluster* tidak perlu mempunyai nilai yang besar untuk menentukan struktur grup yang terikat dari *dataset*.

Nilai juga selalu sama dengan nilai konstan (contoh *N*) dalam pendekatan perankingan berbasis partisi,untuk alasan ini nilai itu dapat menghindarkan dari masalah bias. Maka dari itu,dalam pendekatan berbasis dual partisi,dua fungsi yang membantu dengan derajat normal yang berbeda dari bagian agregasi di tuliskan sebagai berikut :

. . . . . . . . . . . . . (4)

. . . . . . . . . . . . . (5)

Dengan normalisasi pada *J*ss-2 dan pada *Jss-1* sehingga didapatkan nilai konstan untuk bagian agregasi dalam persamaan Seperti dalam pendekatan berbasis pasrtisi-perankingan. Normalisasi ini digunakan untuk menghindari bias dan juga mengurangi kemungkinan komputasi yang berlebihan.

Pengetahuan terdahulu digunakan dalam proses pengelompokan. Pengetahuan ini diberikan dua set aturan berpasangan. Set satu berisi *must-link* (ML)atau harus terhubung dan yang kedua *can not link* (CNL) tidak dapat terhubung. Diasumsikan setiap dokumen mempunyai label virtual berupa variabel kelompok. Label ini terdiri dari dua nilai,satu kelompok pemberian pengguna,yang kedua berdasar kenyataan jika informasi tertentu tersedia. Setiap dokumen dengan label ML mempunyai kemiripan isi antar dokumen tersebut. Sedangkan dokumen dengan label CNL mempunyai ketidak miripan isi antar dokumen tersebut. Aturan ML merepresentasikan relasi yang ekuivalen. Maka,memungkinkan mendapatkan koleksi *transitive closure* dari set ML,dimana setiap dokumen yang berpasangan mempunyai label virtual yang sama. Setiap dokumen didalam aturan berpasangan diberikan derajat yang tinggi terhadap *c* sesuai label virtualnya,dan derajat rendah terhadap kelompok yang lain. Dalam proses klaterisasi bagian seharusnya mempunyai nilai maksimal jika *xi* dan *xk* mempunyai label virtual yang sama dan mempunyai nilai minimal *xi* dan *xk* mempunyai label vistual yang tidak sama. Sehingga kombinasi bagian teramati dalam fungsi objektif menjadi dan mempunyai nilai yang seharusnya maksimal. *Td* adalah faktor nilai bobot yang mengontrol pentingnya pengetahuan terdahulu dari domain dokumen dibandingkan dengan keseluruhan *dataset*. SS-HFCR memastikan setiap dokumen akan mendapatkan distribusi keanggotaan *fuzzy* dan pelanggaran terhadap aturan berpasangan menjadi minimal saat di akhir proses pengelompokkan. *Tu* dan *Tv* digunakan untuk menyesuaikan level keanggotaan *fuzzy* dari dokumen dan kata.

Permasalahan yang dihadapi sekarang adalah memaksimalkan persamaan (4) dan (5). Permasalahan ini dapat diatasi dengan mencari nilai optimal dai **U** dan **V** dengan aturan persamaan (2) dan (3). Karena *u* dan *v* adalah variabel kontinu,dapat menggunakan metode *Lagrange multipliers* dengan dengan urutan pertama kondisi yang diperlukan untuk menurunkan aturan pembaharuan untuk *u* dan *v.* Fungsi *Langgrangian* untuk (4) dan (5) terbentuk seperti :

. . . . . . . . . .. .. .. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . (6)

.. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . (7)

Dimana dan adalah pengali *Lagrange* sesuai dengan aturan (2) dan (3). Untuk memaksimalkan fungsi *Lagrangian L1*,dapat menurunkan secara parsial *L1* untuk *uci* .D engan menghitung untuk *uci* dan untuk *vcj* .

. . . . (8)

. . . . . . . .. . . . . . . . . . . .. . . . . . . .. . . . . (9)

# RINGKASAN ISI TUGAS AKHIR

Secara garis besar, proses yang dilakukan pada tugas akhir ini ditunjukkan pada Gambar 1



Gambar 1.Bagan proses sistem

1. **Dataset Dokumen Teks**

*Dataset* dokumen teks yang digunakan dalam tugas akhir ini didapat dari [http://www.cs.umb.edu/~smimarog/textmining/*dataset*s/](http://www.cs.umb.edu/~smimarog/textmining/datasets/) . Data yang digunakan ada tiga yaitu *20 Newsgroup,Reuters 21578,*dan *Webkb* dengan rincian sebagai berikut.

* Data *20 Newsgroups* ,data ini terdiri dari artikel yang terdiri dari 20 grup berita. Selanjutnya data ini dijelaskan dalam tabel sebagai berikut:

|  |  |  |  |
| --- | --- | --- | --- |
| ***20 Newsgroups*** | | | |
| **Kelas** | **Jumlah Data Pelatihan** | **Jumlah Data Pengujian** | **Jumlah Keseluruhan** |
| *alt.atheism* | 480 | 319 | 799 |
| *comp graphics* | 584 | 389 | 973 |
| *comp.os.ms-windows.Misc* | 572 | 394 | 966 |
| *comp.sys.ibm.pc.hardware* | 590 | 392 | 982 |
| *comp.sys.mac.hardware* | 578 | 385 | 963 |
| *comp.windows.x* | 593 | 392 | 985 |
| *misc.forsale* | 585 | 390 | 975 |
| *rec.autos* | 594 | 395 | 989 |
| *rec.motorcycles* | 598 | 398 | 996 |
| *rec.sport.baseball* | 597 | 397 | 994 |
| *rec.sport.hockey* | 600 | 399 | 999 |
| *sci.crypt* | 595 | 396 | 991 |
| *sci.electronics* | 591 | 393 | 984 |
| *sci.med* | 594 | 396 | 990 |
| *sci.space* | 593 | 394 | 987 |
| *soc.religion.christian* | 598 | 398 | 996 |
| *talk.politics.guns* | 545 | 364 | 909 |
| *talk.politics.mideast* | 564 | 376 | 940 |
| *talk.politics.misc* | 465 | 310 | 775 |
| *talk.religion.misc* | 377 | 251 | 628 |
| Jumlah | 11293 | 7528 | 18821 |

* Data *Reuters 21578* akan dibagi menjadi dua,yaitu R8 untuk data dengan 8 kelas dan R52 untuk data dengan kelas berjumlah 52.

Untuk R8 dijelaskan dalam tabel sebagai berikut:

|  |  |  |  |
| --- | --- | --- | --- |
| **R8** | | | |
| **Kelas** | **Jumlah Data Pelatihan** | **Jumlah Data Pengujian** | **Jumlah Keseluruhan** |
| *acq* | 1596 | 696 | 2292 |
| *crude* | 253 | 121 | 374 |
| *earn* | 2840 | 1083 | 3923 |
| *grain* | 41 | 10 | 51 |
| *interest* | 190 | 81 | 271 |
| *money-fx* | 206 | 87 | 293 |
| *ship* | 108 | 36 | 144 |
| *trade* | 251 | 75 | 326 |
| Jumlah | 5485 | 2189 | 7674 |

* Untuk R52 dijelaskan dalam tabel sebagai berikut:

|  |  |  |  |
| --- | --- | --- | --- |
| **R52** | | | |
| **Kelas** | **Jumlah Data Pelatihan** | **Jumlah Data Pengujian** | **Jumlah Keseluruhan** |
| *acq* | 1596 | 696 | 2292 |
| *alum* | 31 | 19 | 50 |
| *bop* | 22 | 9 | 31 |
| *carcass* | 6 | 5 | 11 |
| *cocoa* | 46 | 15 | 61 |
| *coffee* | 90 | 22 | 112 |
| *copper* | 31 | 13 | 44 |
| *cotton* | 15 | 9 | 24 |
| *cpi* | 54 | 17 | 71 |
| *cpu* | 3 | 1 | 4 |
| *crude* | 253 | 121 | 374 |
| *dlr* | 3 | 3 | 6 |
| *earn* | 2840 | 1083 | 3923 |
| *fuel* | 4 | 7 | 11 |
| *gas* | 10 | 8 | 18 |
| *gnp* | 58 | 15 | 73 |
| *gold* | 70 | 20 | 90 |
| *grain* | 41 | 10 | 51 |
| *heat* | 6 | 4 | 10 |
| *housing* | 15 | 2 | 17 |
| *income* | 7 | 4 | 11 |
| *instal-debt* | 5 | 1 | 6 |
| *interest* | 190 | 81 | 271 |
| *ipi* | 33 | 11 | 44 |
| *iron-steel* | 26 | 12 | 38 |
| *jet* | 2 | 1 | 3 |
| *jobs* | 37 | 12 | 49 |
| *lead* | 4 | 4 | 8 |
| *lei* | 11 | 3 | 14 |
| *livestock* | 13 | 5 | 18 |
| *lumber* | 7 | 4 | 11 |
| *meal-feed* | 6 | 1 | 7 |
| *money-fx* | 206 | 87 | 293 |
| *money-supply* | 123 | 28 | 151 |
| *nat-gas* | 24 | 12 | 36 |
| *nickel* | 3 | 1 | 4 |
| *orange* | 13 | 9 | 22 |
| *pet-chem* | 13 | 6 | 19 |
| *platinum* | 1 | 2 | 3 |
| *potato* | 2 | 3 | 5 |
| *reserves* | 37 | 12 | 49 |
| *retail* | 19 | 1 | 20 |
| *rubber* | 31 | 9 | 40 |
| *ship* | 108 | 36 | 144 |
| *strategic-metal* | 9 | 6 | 15 |
| *sugar* | 97 | 25 | 122 |
| *tea* | 2 | 3 | 5 |
| *tin* | 17 | 10 | 27 |
| *trade* | 251 | 75 | 326 |
| *veg-oil* | 19 | 11 | 30 |
| *wpi* | 14 | 9 | 23 |
| *zinc* | 8 | 5 | 13 |
| Total | 6532 | 2568 | 9100 |

* Data *Webkb* dijelaskan dalam tabel sebagai berikut:

|  |  |  |  |
| --- | --- | --- | --- |
| **WebKB** | | | |
| **Kelas** | **Jumlah Data Pelatihan** | **Jumlah Data Pengujian** | **Jumlah Keseluruhan** |
| *project* | 336 | 168 | 504 |
| *course* | 620 | 310 | 930 |
| *faculty* | 750 | 374 | 1124 |
| *student* | 1097 | 544 | 1641 |
| *Jumlah* | 2803 | 1396 | 4199 |

1. **Praproses Dataset**

Praproses *dataset* dilakukan untuk mengolah data agar dapat diproses dengan baik selama implementasi metode. Praproses dilakukan dengan cara membentuk *Vector Space Model (VSM)*atau matriks *tf-idf*.

1. **Pengelompokan Dokumen Menggunakan Metode SS-HFCR**

Pengelompokan dataset menggunakan metode yang telah diajukan yaitu *SS-HFCR*. Implementasi menggunakan bahasa pemrograman python. Librari scipy,numpy dan sklearn. Gambaran besar dari metode sebagai berikut:

***The* SS-HFCR *algorithm*.**

***Input***:*Dataset* **D**, *number of clusters* C, *Constraint sets*: ***ML&CNL***

***Output***:*Document membership matrix*:**U**,*Word membership matrix*: **V**.

***Method***:

1. *Set weighting factor* Tu,Tv,Td, *stopping threshold* , *maximum iteration number*
2. *Manually adjust the initial* *uci* *for the documents existed in* ***ML&CNL*** *set to obey all the constraints, then randomly assign the initial* *uci* *for the other documents*.
3. *REPEAT*

3. 1 *Update vcj with Eq*. (9);

3. 2 *Update* *uci* *with* *Eq*. (8)

3. 3

*UNTIL*

1. **Evaluasi Hasil Pengelompokkan**

Akurasi,kestabilan dan efesiensi adalah kriteria yang penting dalam evaluasi performa algoritma pengelompokkan. Maka dari itu,performa SS-HFCR dinilai dalam kriteria itu. Evaluasi yang akan dilakukan sebagai berikut:

1. Untuk akurasi digunakan *Accuracy* dan *normalized mutual information* (NMI).

*Accuracy = ...........................(10)*

*Accuracy* merupakan pendekatan seberapa akurat metode pengelompokan memberikan label kelompok *ji* terhadap kenyataan kelompok *ci*. NMI digunakan untuk mengantisipasi jika jumlah kelompok yang diberikan algoritma tidka sesuai jumlah kelompok yang diberikan oleh pengguna.

NMI= ......................(11)

1. Untuk kestabilan dinilai seberapa sensitif algoritma terhadap variasi aturan berpasangan. Dengan kata lain kestabilan menguji apakah hasil pengelompokan adalah “*most stable*” terhadap pengambilan secara acak pasangan dokumen sebagai aturan. Pengukuran kestabilan dilakukan dengan menghitung deviasi dari 20 percobaan setiap dataset untuk setiap aturan. Deviasi yang rendah menunjukkan bahwa SS-HFCR dapat menghasilkan pengelompokan yang stabil untuk dengan aturan berpasangan yang berbeda.
2. Untuk efesiensi dinilai berdasar kompleksitas waktu dan *run time*

# METODOLOGI

## Penyusunan proposal tugas akhir

Proposal tugas akhir ditulis untuk mengajukan ide atas pengerjaan tugas akhir. Proposal tugas akhir juga mengandung proyeksi hasil dari ide tugas akhir yang diajukan.

## Studi literatur

Pada tahap ini dilakukan pengumpulan *dataset* dari [http://www.cs.umb.edu/~smimarog/textmining/*dataset*s/](http://www.cs.umb.edu/~smimarog/textmining/datasets/) dan juga pemahaman literatur yang berhubungan dengan metode *SS-HFCR*. Literatur yang digunakan dalam pengerjaan tugas akhir ini sebagian besar berasal dari internet berupa artikel , materi kuliah , serta beberapa buku referensi.

## Implementasi Metode

Implementasi merupakan tahap untuk membangun sistem tersebut. Terdapat algortima *SS-HFCR* yang akan diimplementasikan. Implementasi menggunakan python dengan librari *sklearn*,*numpy* dan *scipy*. Sedangkan preproses dilakukan dengan bahasa pemrograman java dan librari weka.

## Pengujian dan Evaluasi

Akurasi,kestabilan dan efesiensi adalah kriteria yang penting dalam evaluasi performa algoritma pengelompokkan. Maka dari itu,performa SS-HFCRdinilai dalam kriteria itu.

## Penyusunan Buku Tugas Akhir

Pada tahap ini disusun laporan tugas akhir sebagai dokumentasi pelaksanaan tugas akhir, yang mencakup seluruh konsep, teori, implementasi, serta hasil yang telah dikerjakan. Laporan tugas akhir ini akan dibagi menjadi beberapa bab sebagai berikut :

1. Pendahuluan
   1. Latar Belakang
   2. Rumusan Masalah
   3. Batasan Tugas Akhir
   4. Tujuan
   5. Metodologi
   6. Sistematika Penulisan
2. Tinjauan Pustaka
3. Desain dan Implementasi
4. Pengujian dan Evaluasi
5. Kesimpulan dan Saran
6. Daftar Pustaka

# JADWAL KEGIATAN

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tahapan | Tahun | | | | | | | | | | | | | | | | |
| September | | | | Oktober | | | | Nopember | | | | Desember | | | | |
| Penyusunan Proposal |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Studi Literatur |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Implementasi Metode |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Pengujian dan evaluasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Penyusunan buku |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

# DAFTAR PUSTAKA

|  |  |
| --- | --- |
| [1] | (2013, Apr.) Wikipedia. [Online]. <http://id.wikipedia.org/wiki/Dokumen> |
| [2] | K. P. Murphy, *Machine Learning: a Probabilistic Perspective*. Cambridge,Massachusetts, London,England: The MIT Press, 2013. |
| [3] | Y. Yan, L. Cheb, and W.-C. Tjhi, "Fuzzy semi-supervised co-clustering for text documents," *Fuzzy Sets And System*, p. 74, 2013. |
| [4] | W.-C. Tjhi and L. Chen, "A heuristic fuzzy co-clustering algorithm for categorization of high dimensional data," *Fuzzy Sets And System*, p. 371, 2008. |
| [5] | (2007) Some Text Dataset. [Online]. <http://www.cs.umb.edu/~smimarog/textmining/datasets/> |

1. www. cs. umb. edu/~smimarog/textmining/*dataset*s/ [↑](#footnote-ref-1)