Indexação automática

CBD/ECA – Indexação: teoria e prática

- Indexação documentária
  - Identificar pela análise dos documentos, os seus assuntos
  - extrair os conceitos que indicam o seu conteúdo
  - traduzir os conceitos em descritores, utilizando um vocabulário controlado (linguagem documentária), visando à sua recuperação posterior.

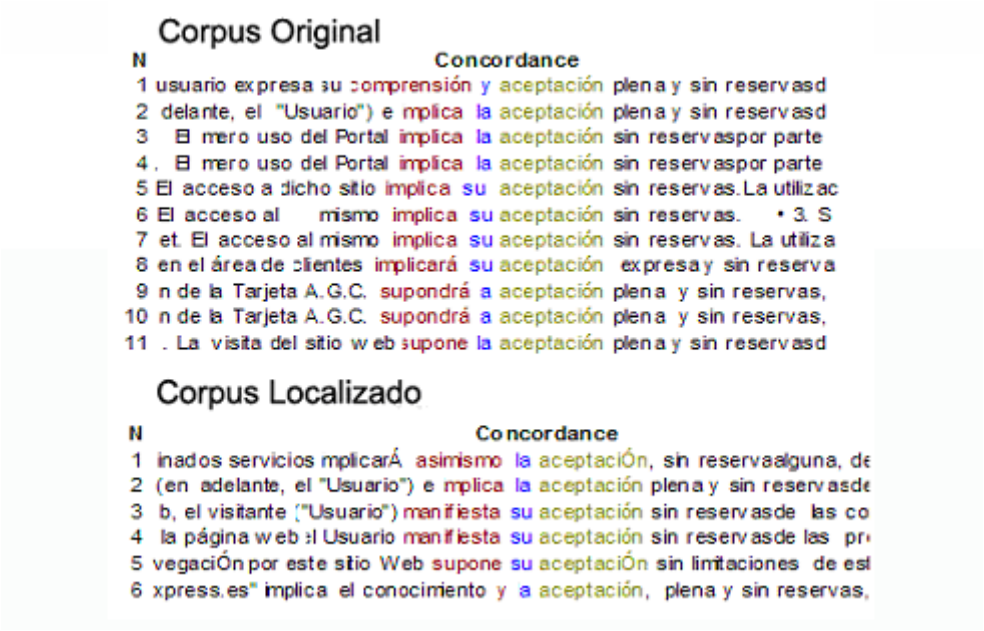
- **Indexação manual**
  - Denominada também de indexação intelectual ou humana, é aquela realizada pelo bibliotecário (profissional da informação, indexador).
  - **Etapas**
    - leitura documentária para a identificação do conteúdo informacional do documento
    - seleção dos conceitos pertinentes
    - representação (“tradução”) dos conceitos selecionados em descritores da linguagem documentária adotada pelo Sistema de Informação.

- **Indexação automática**
  - indexação por computador (*computer indexing),*
  - indexação mecânica *(mechanical indexing)*,
  - programa de indexação (*indexing program)*,
  - indexação automática (*automatic indexing),*

- **Tipos de indexação automática**
  - programas que auxiliam o processo de armazenamento de termos de indexação, obtidos de modo intelectual, mais conhecido como armazenamento de termos de indexação assistido por computador;
  - sistemas que analisam os documentos de modo automático, mas os termos de indexação propostos são validados e editados por um profissional (indexação semi-automática);
  - programas sem nenhum tipo de validação, isto é, os termos propostos são armazenados diretamente como descritores do dito documento (indexação automática).

- **Histórico**
  - Luhn (1958 e 1959): critérios estatísticos de ocorrência/freqüência.
    - o índice KWIC (key word in context), as palavras do título que servem de entradas no índice são identificadas automaticamente a partir da eliminação das palavras não significativas, por comparação com uma lista de palavras vazias de significado, estabelecida previamente.
  - Schank & Abelson (1977) e Lehnert (1984) : geração de resumos automáticos, utilizados como mecanismos de validação da compreensão de textos de natureza diversa:
    - textos jornalísticos, mensagens de telex, narrativas
  - Gardin (1977): aproximação da Análise Documentária da Lingüística, via tradução automática

- Imagem de um índice kwic automático utilizado em localização e tradução
- Fonte: http://www.fti.uab.es/tradumatica/revista/num7/articles/05/05art.htm



Corpus Original

N	Concordance
1	usuario expresa su comprensión y aceptación plena y sin reservasd
2	delante, el "Usuario") e mplica la aceptación plena y sin reservasd
3	El mero uso del Portal implica la aceptación sin reservaspor parte
4	. El mero uso del Portal implica la aceptación sin reservaspor parte
5	El acceso a dicho sitio implica su aceptación sin reservas.La utilizac
6	El acceso al mismo implica su aceptación sin reservas. • 3. S
7	et. El acceso al mismo implica su aceptación sin reservas. La utiliza
8	en el área de clientes implicará su aceptación expresa y sin reserva
9	n de la Tarjeta A.G.C. supondrá a aceptación plena y sin reservas,
10	n de la Tarjeta A.G.C. supondrá a aceptación plena y sin reservas,
11	. La visita del sitio w eb supone la aceptación plena y sin reservasd

Corpus Localizado

N	Concordance
1	inados servicios mplicará asimismo la aceptaciÓn, sin reservaalguna, de
2	(en adelante, el "Usuario") e mplica la aceptación plena y sin reservasde
3	b, el visitante ("Usuario") manifiesta su aceptación sin reservasde las co
4	la página w eb :l Usuario manifiesta su aceptación sin reservasde las pr
5	vegaciÓn por este sitio Web supone su aceptaciÓn sin limitaciones de est
6	xpress.es" implica el conocimiento y a aceptación, plena y sin reservas,

- Abordagem lingüístico-computacional:

  – através de processos sucessivos de análise léxica, sintática e semântica das orações do texto, com ajuda de dicionários de palavras vazias e significativas, para identificação dos descritores. Pela sua contribuição teórica.

  – os trabalhos de Gardin e sua equipe no desenvolvimento do SYNTOL merecem uma menção especial (syntagmatic organization language), uma nova e original linguagem documentária com aplicação à identificação de descritores e à recuperação da informação.

- **Experiências no Brasil**

  – Robredo (1991): propôs um sistema para gerar termos indexadores a partir da análise automática de títulos e resumos de textos **(AUTOINDEX),** onde os termos candidatos a descritores são selecionados pela comparação do texto com 2 dicionários:

    - "palavras vazias invariáveis" (preposições, conjunções, advérbios, etc.)
    - raízes de palavras tidas como "não significativas na área de conhecimento em questão

  – O conjunto extraído é, em seguida, submetido a tratamentos estatísticos (freqüência) para a determinação da relevância de cada unidade do texto.

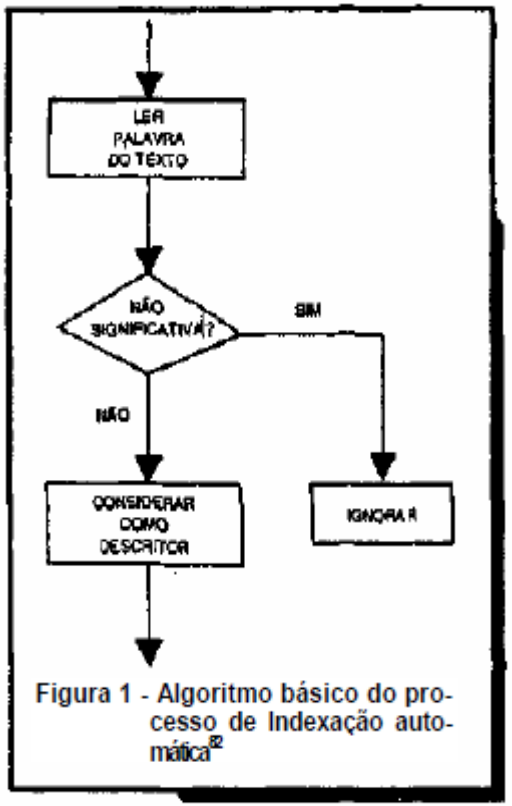


Figura 1 - Algoritmo básico do processo de Indexação automática[82]

A	CENTRAL	EFICIENTES
ABAIXO	CINCO	EIS
ABRIL	COM	ELA
ACAO	COMO	ELAS
ACERCA	COMPARADA	ELE
ACIMA	CONTRA	ELES
.	.	EM
.	.	EMBORA
.	.	ENFIM
AFORA	DA	ENQUANTO
AGORA	DAÍ	ENTAO
AGOSTO	DALI	ENTRE
AI	DAQUI	.
AINDA	DAS	.
ALEM	DE	.
AMANHA	DELA	ESTA
AMBAS	DELAS	ESTAS
AMBOS	DELE	ESTE
ANO	DELES	ESTES
ANOS	DEMAIS	EXCETO
.	.	.
.	.	.
.	.	.
BOA	DO	FOR
BOAS	DOIS	FORA
BOM	DOS	.
BONS	DOZE	.
.	DUAS	.
.	DURANTE	FRENTE
.	DUZENTOS	.
CA	.	.
CADA	.	.
CEM	.	UM
CENTO	E	UMA
.	.	.
.	.	.
.	.	.

Figura 2- Fragmento do dicionário de palavras vazias de significado ou antidicionário (wordfixed)

ABAL	ASPEC	COINCID
ABANDON	ASPIRA	COISA
ABERT	ASSEGUR	COLIG
ABORD	ASSESS	COLOC
ABR	ASSIST	COME
ABSOL	ASSOCIAD	COMPET
.	.	COMPLET
.	.	COMPREE
.	.	COMPRO
ADMIR	BARAT	COMUM
ADMIS	BASEA	CONCED
ADMIT	BASICA	CONCI
ADOT	BASICO	.
ADVER	BASTA	.
AFET	BATE	.
FIRM	BENEF	CONTID
AFLIC	BONI	CONTIN
AFLIG	BRANC	CONTOR
AFRONT	BRILH	CONTRAD
AJUD	BRINC	CONTRAR
.	.	.
.	.	.
.	.	.
ALGUM	CAMINH	CONTRIB
ALGUN	CAMP	.
.	.	.
.	.	.
.	.	CUMPR
APAREC	CHEF	CURIOS
APAREN	CHES	CURT
.	.	.
.	.	.
.	.	.

Figura 3-Fragmento do antidicionário de raízes de palavras não significativas (wordroots)

Parêntese quadrado (abre)	[
Parêntese quadrado (fecha)	]
Parêntese (abre)	(
Parêntese (fecha)	)
Ponto	.
Dois Pontos	:
Vírgula	,
Ponto e Vírgula	;
Asterisco	*
Apóstrofo	'
Exclamação	!
Interrogação	?
Barra diagonal	/
Aspas	"
Espaço	

Figura 4 – Delimitadores de palavras

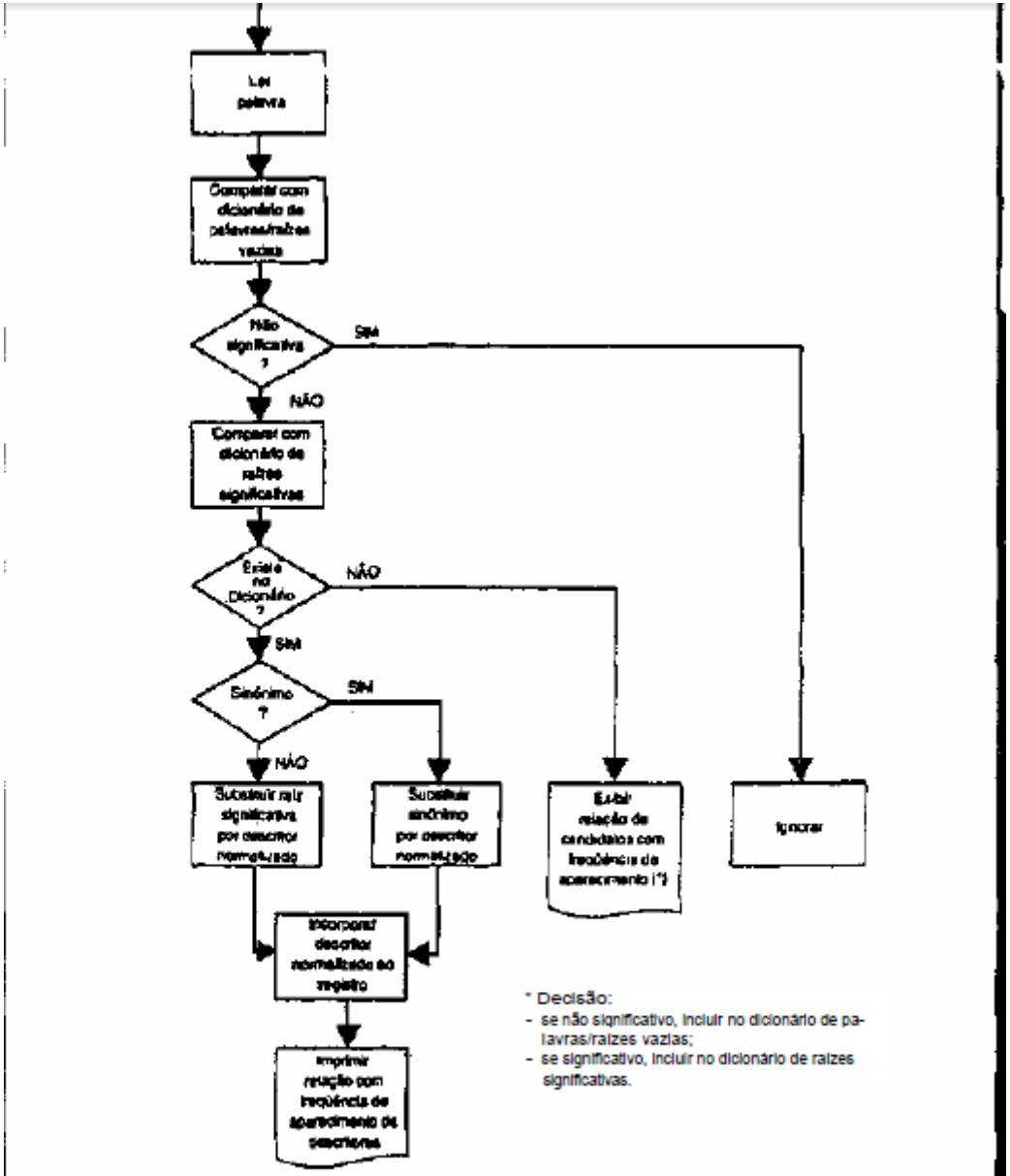


Figura 5 - Algoritmo simplificado do processo de indexação automática


```
a)
CI00119
FREUND, G. E.
UNIVERSIDADE DE SÃO PAULO.
"ANALISE ESTRUTURAL PARA AUMENTAR A EFICIENCIA DE PESQUISAS
ONLINE".
CIENCIA DA INFORMACAO. V.11, N.1, P.19-26. 1982.
PROPOE UMA TECNICA BASEADA NA ANALISE SINTATICA DE TERMOS CO
MO OBJETIVO DE IDENTIFICAR PALAVRAS SEMANTICAMENTE RELACIONA
DAS PARA SOLUCAO DE PROBLEMAS, COMO O DO TRUNCAMENTO ARBITRA
RIO.

b)
ANALISE-ESTRUTURAL PARA AUMENTAR A EFICIENCIA DE PESQUISAS-O
NLINE.
PROPOE UMA TECNICA BASEADA NA ANALISE-SINTATICA DE TERMOS CO
M O OBJETIVO DE IDENTIFICAR PALAVRAS-SEMANTICAMENTE-RELACION
ADAS PARA SOLUCAO DE PROBLEMAS, COMO O DE TRUNCAMENTO-ARBITR
ARIO.

c)
ANALISE-ESTRUTURAL      PALAVRAS-SEMANTICAMENTE-RELACIONADAS
ANALISE                 palavras (PALAVRAS)
estrutural (ESTRUTURA)  semanticamente (SEMANTICA)
EFICIENCIA              TRUNCAMENTO-ARBITRARIO
PESQUISAS-ONLINE        TRUNCAMENTO
pesquisas (PESQUISA)
ONLINE
TECNICA
ANALISE-SINTATICA
sintatica (SINTAXE)
termos (TERMO)
```

Figura 6 - Processo de indexação automática a partir dos títulos e resumos dos registros documentários

FONTE	CI00050
FONTE	CI00126
FONTES	CI00064
FONTES-DE-INFORMAÇÃO	CI00047
FORMAÇÃO	CI00099
FORMAÇÃO-PROFISSIONAL	CI00019
FORMATO	CI00011
FÓRMULA-DE-TRANSIÇÃO	CI00024
FORNECEDORES-DA-INFORMAÇÃO	CI00144
FORNECIMENTO-DE-LIVROS	CI00005
FORTALECIMENTO	CI00138
FRANCA	CI00061
FRASES	CI00025
FRENTE-DE-PESQUISA	CI00016
FRENTE-DE-PESQUISA	CI00044
FRENTE-DE-PESQUISA	CI00094
FREQUENCIA	CI00029
FREQUENCIA	CI00032
FUTUROLOGIA	CI00019
GATEKEEPERS	CI00109
GEOLOGICA	CI00017
GEORGES-ANDERLA	CI00071
GERAÇÃO	CI00114

Figura 7 - Fragmento do índice temático impresso, obtido mediante o processo de indexação automática, a partir do título e do resumo dos artigos publicados na revista *Ciência da Informação*[83]


```
UNIVERSIDADE-FEDERAL-DA-BAHIA
ESCOLA-DE-BIBLIOTECONOMIA-E-DOCUMENTACAO
SALVADOR, 03-ABR-89

ILMO. SR.
PROF. JAIME-ROBREDO
BRASILIA, DF

PREZADO PROFESSOR,
DE CONFORMIDADE COM OS ENTENDIMENTOS POR TELEFONE, MANTIDOS COM
V.SA., CONFIRMAMOS O PERIODO DE 03-07 JUL PARA A REALIZACAO DO
CURSO DE INTRODUCAO AOS PROCESSOS DE INDEXACAO-AUTOMATICA DE TE
XTOS. INFORMAMOS, OUTROSSIM, QUE O REFERIDO CURSO JÁ SE ENCONTR
A APROVADO PELA COORDENACAO-DE-EXTENSAO DA UFBA, DE CONFORMIDAD
E COM O PROGRAMA E CARGA-HORARIA DE 40 HORAS, ESTABELECIDOS POR
V.SA.
ESTAMOS ANCAMINHANDO, NA PRESENTE DATA, OFICIO DO DR. MARCOS-FO
RMIGA, DO INEP, VISANDO A CONFIRMACAO DA PASSAGEM-AEREA BRASILI
A-SALVADOR-BRASILIA, A SER UTILIZADA POR V.SA.  PARA REALIZACAO
DO MESMO CURSO.

ATENCIOSAMENTE.
MARGARIDA-PINTO-DE-OLIVEIRA
COODENADORA-DO-CURSO
```

Figura 8 - Texto de uma carta, após editoração

UNIVERSIDADE-FEDERAL-DA-BAHIA
UNIVERSIDADE
FEDERAL
BAHIA
ESCOLA-DE-BIBLIOTECONOMIA-E-DOCUMENTACAO
BIBLIOTECONOMIA
DOCUMENTACAO
SALVADOR
03-ABR-89
JAIME-ROBREDO
BRASILIA
DF
entendimento (ENTENDIMENTO)
TELEFONE
confirmamos (CONFIRMACAO)
03-07-JUL
CURSO
INTRODUCAO
PROCESSOS
INDEXACAO-AUTOMATICA
INDEXACAO
automatica (AUTOMACAO)
textos (TEXTO)
informamos (INFORMACAO)
aprovado (APROVACAO)
COORDENACAO-DE-EXTENSAO
COORDENACAO
EXTENSAO
UFBA
PROGRAMA
CARGA-HORARIA
40-HORAS
encaminhamos (ENCAMINHAMENTO)
DATA
OFICIO
MARCOS-FORMIGA
INEP
PASSAGEM-AEREA
BRASILIA-SALVADOR-BRASILIA
MARGARIDA-PINTO-DE-OLIVEIRA
COORDENADORA DO CURSO

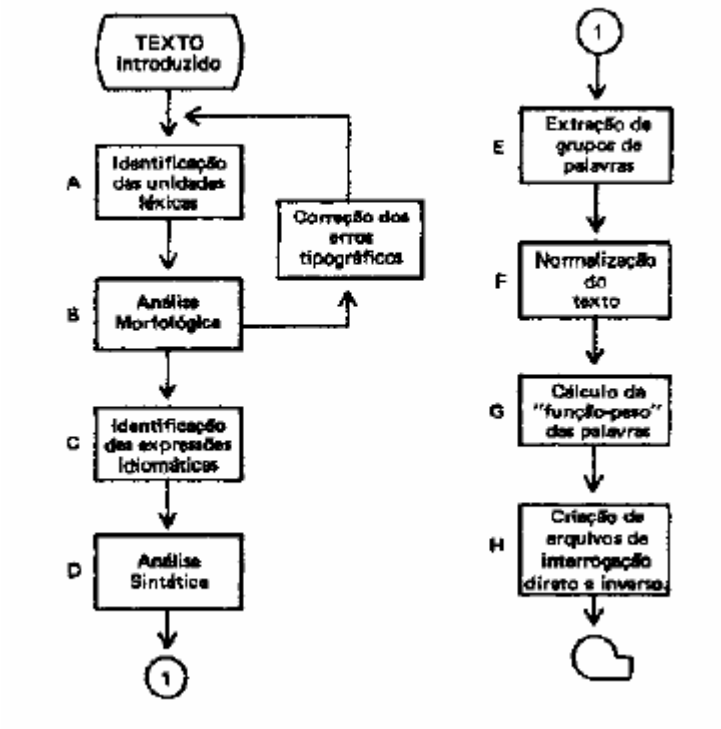
Figura 9 - Descritores/palavras-chave e/ou candidatos extraídos do texto da carta a que se refere a figura 8

- **Experiências no Brasil**

- SPIRIT (1983)

  – desenvolvido para o idioma francês por uma equipe de pesquisadores do CNRS e adaptado ao português.
  – processamento de documentos em linguagem natural com o auxílio de métodos linguísticos combinados com métodos estatísticos permitindo uma indexação ponderada.

3.1 Tratamento de um Corpus



- **Experiências no Brasil**
  - Kuramoto (1996): propôs um modelo de indexação automática, de base lingüística, centrada na identificação de sintagmas nominais.

- **Indexação automática**
  - Ocorre quando um programa de computador, adotando critérios de freqüência, extrai palavras, expressões ou radicais de palavras do texto para representar o seu conteúdo como um todo. (Santos e Ribeiro, 2003)
    - Indexação por extração automática
    - Indexação por atribuição automática

- **Indexação por extração automática**

  – Palavras ou expressões do texto são extraídas, por um software, e utilizadas para representar o conteúdo do texto, como um todo
  – Critérios de seleção: freqüência, posição e contexto.

  – O software
    - conta as palavras do texto, compara-os com uma lista de palavras proibidas;
    - elimina palavras não-significativas (artigos, preposições, conjunções etc.)
    - ordena as palavras segundo a freqüência de sua ocorrência.

  – Palavras com **>** número de ocorrências = Descritores

  (LANCASTER, 2004, p.286)

- **Indexação por atribuição automática**

  – Para cada termo a ser atribuído um “perfil” de palavras ou expressões que costumam ocorrer freqüentemente nos documentos.

  – O software compara o perfil dos termos do documento com o perfil dos termos freqüentemente atribuídos por indexador humano

  – O software atribui um termo ao documento, sempre que o perfil do documento coincide com o perfil dos termos.

(LANCASTER, 2004, p.289).

- **Indexação semi-automática**

  – Combinação de indexação manual e automática.

  – ETAPAS

    - inicialmente, o sistema faz indexação automática levando em conta as ocorrências das palavras mais freqüentes no texto.

    - em um segundo momento, indexador humano refina a lista dos descritores propostos pelo sistema, fazendo ajustes e/ou complementações necessárias.

(PINTO, 2001, p.227)

- **Softwares para indexação**
  - **Objetivo:** descentralizar a produção de informações documentárias como também distribuí-las de forma extensiva e rápida.
  - **Sistema de Indización Automático (SISA)**
    - Ferramenta para assistir a indexação.
    - Desenvolvido em 1997 por Isidoro Gil Leiva (Universidade de Múrcia , Espanha)

- **Requisitos do SISA**
  - Todos os textos a serem indexados devem estar no mesmo diretório, em formato txt, marcados com os seguintes parâmetros :
    - #CTI# e #FTI# para identificar o título;
    - #CR# e #FR# para identificar o resumo;
    - #CTE# e #FTE# para identificar o texto do artigo.
    - cada linha do texto não deve ter mais do que 100 palavras

- **Default do SISA**
  - ***Vocabulario.txt*** = lista de descritores em Ciência da Informação
  - ***TG.txt*** = lista de termos de Ciência da Informação retirados de
    - *Tesauro en documentacíon e informacíon";*
    - *Tesauro de la UNESCO*
    - *"Vocabulário controlado en bibliotecologia, Ciência de la información y temas afines"*
  - ***Vacias.txt*** = lista de palavras vazias em espanhol

- **Funcionamento do SISA :** Indexa simultaneamente até 10 textos

  - **Etapas:**
    - Horizontalização do texto
    - Eliminação de palavras vazias (palavras com funções apenas gramaticais, como por exemplo os conectivos, artigos, pronomes, etc)
    - Comparação dos termos retidos com os descritores do vocabulário controlado do sistema
    - Segmentos considerados: título, resumo e corpo do texto.
    - Apresentação de todos os termos candidatos à indexação
      - termos do vocabulário controlado
      - palavras que não estão contidas no vocabulário controlado do sistema, mas que apresentam várias ocorrências no título, no resumo e no texto completo.
      - permite ao indexador tomar a decisão de utilizá-las ou não na indexação final
      - permite exportar os resultados para um arquivo denominado *Resultados.*

- **Para indexação em português**
  - Arquivos inseridos na pasta ***Config*** do programa:
    - ***Vocabulário.txt***: assuntos principais do “Vocabulário Controlado USP” (VOCAUSP)
    - ***TG.txt***: lista dos assuntos principais relacionados hierarquicamente
    - ***Vacias.txt***: lista de palavras vazias em português

- Exercício em sala: material disponibilizado no Stoa Moodle

- Trabalho final da disciplina: indexação manual e automática (comparação com subsídio teórico).
  - Selecionar artigos sobre o mesmo assunto e/ou revista (5 por aluno)
  - Preparar os artigos:
    - salvar em formato txt;
    - inserir etiquetas de marcação
    - reunir os arquivos em um mesmo diretório.
  - Indexar com o SISA
  - Exportar resultados
  - Abrir arquivo resultados.txt, inserir o comando **#TODOS#** e salvar arquivo.

Referências:

ANDREEWISKI, A., Ruas, V. Indexação automática baseada em métodos linguisticos e estatísticos e sua aplicabilidade em língua portuguesa. **Ciência da Informação**, v.12, n.1, 1983.

GIL-LEIVA, I. **La automatizacion de la indizacion de documentos**. Gijón: Ediciones Trea, 1999.

KURAMOTO, H. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação**, Brasília, v. 25, n. 2, 1996.

LANCASTER, F. W. **Indexação e resumos: teoria e prática**. Brasília: Briquet de Lemos/Livros, 1993, 1997, 2004

ROBREDO, J. Indexação automática de texto: uma abordagem otimizada e simples. **Ciência da Informação**, v. 20, n. 2, 1991.

VIEIRA, S. B. Indexação automática e manual: revisão de literatura. **Ciência da Informação**, v.17, n.1, 1988.