Semantic API Specification

Rev. 2.0

1. Introduction

The following API methods are described in this document:

- ExtractConcepts
- ExtractCategories
- CreateSummary

The set of their parameters and the format of results represent most common scenarios of usage that the authors have thought of. However, a number of variations of these methods are available as well. You may request a variation that fits your needs. In many cases, it will be made available to you at no additional charge.

As you can see from the document, there are a lot of parameters allowing you to tweak the type and format of results. We recommend that you use the examples as your starting point in familiarizing yourself with using the API and finding the parameters that are right for you.

The API supports SOAP and REST protocols (HTTP GET or POST).

2. ExtractConcepts method

If you use SOAP, you can review the schema at: http://www.sensebot.net/svc/extconcone.asmx

Description: this method takes a document HTTP URL as input and returns a list of key semantic concepts extracted from the document. The URL should refer to a document or page in HTML or TXT format.

Parameters (all mandatory):

Name	Type	Description	Comments
username	String	API key	The API key that is created for you when you sign up for the API
allURLs	string, URL-encoded	URL of the source document or page (starting with http://)	The source has to be in HTML or text format
artClass	String	This parameter is reserved for the future; pass empty string for now.	
artLength	String	This parameter is reserved for the future; pass 0 for now.	
numConcepts	String	Number of concepts to return. Use a whole number between 1 and 6.	
Lang	string	Language of the source document. Available languages: "English", "French", "Spanish",	

		"German".	

Result:

An XML string containing the list of concepts. Each concept can consist of one word or several words. Each concept is followed by colon and number "1", e.g. "computer:1". The concepts are returned in no particular order. The concepts are separated by space character.

EXAMPLE: extract semantic concepts from CNN.com page using HTTP GET:

GET URL:
http://*servername*/svc/extconcone.asmx/ExtractConcepts?userName=12345&numConcepts=3&artClass=&artLength=0&lang=English&allURLs=http%3A//www.cnn.com/

(Substitute *servername* with the actual server domain you are connecting to).

Result returned:

```
<?xml version="1.0" encoding="utf-8"?>
<string xmlns="http://sensebot.net/svc/">CRASH:1 BULL:1 POPULAR:1 PLANE:1
WEATHER:1 VICTIMS:1 </string>
```

3. ExtractCategories method

If you use SOAP, you can review the schema at: http://www.sensebot.net/svc/extconcone.asmx

Description: This method extracts semantic concepts from a webpage and maps them to a set of advertising categories. The method arguments are the same as for the **ExtractConcepts** method described in the documentation. The method URL will start as follows:
http://api.sensebot.net/svc/extconcone.asmx/ExtractCategories?…

The return string will contain a list of categories prefixed by the title CATEGORIES: and separated by semicolons.

EXAMPLE: extract categories from an article using HTTP GET:

GET URL:
http://*servername*/svc/extconcone.asmx/ExtractCategories?userName=12345&numConcepts=3&artClass=&artLength=0&lang=English&allURLs=http%3A//shine.yahoo.com/channel/beauty/the-foolproof-anti-aging-skin-routine-2401689

(Substitute *servername* with the actual server domain you are connecting to).

Result returned:

CATEGORIES: Beauty; Grocery; CONCEPTS: SKIN:1 CREAM:1 EYE:1 MOISTURE:1 BEAUTY:1 ACIDS:1

4. CreateSummary method

If you use SOAP, you can review the schema at: http://www.sensebot.net/svc/contentsvc.asmx

Description: this method takes a set of HTTP URLs as input and returns a multi-document summary of the sources, and/or a list of key semantic concepts. The URLs should refer to documents or pages in HTML or TXT format.

Parameters (all mandatory):

Name	Type	Description	Comments
Username	string	API key	The API key that is created for you when you sign up for the API
allURLs	string, URL-encoded	A string of URLs of the sources, separated by ` (HTML code `)	The number of URLs has to match **numSources**. Supported pages are in HTML or TXT format.
allTitles	string, URL-encoded	A string of titles of the sources, separated by ` (HTML code `)	The number of Titles has to match **numSources**
numSources	string	Number of sources to process	In the standard deployment of the API the value of **numSources** is limited to 10 (can be extended per client needs)
Query	string	Query of the request, used in setting the focus of semantic analysis. Also serves as the title of the result summary.	No Boolean logic is supported in the query
numSentences	string; the value should be greater than "0"	Desired length of the summary, in sentences.	If the summary is not requested (i.e., **bSummary** equals 0), the value of this parameter is irrelevant
lang	string	Language of the sources. Available languages: "English", "French", "Spanish", "German", "Japanese".	
simThreshold	string; the value range is between "0" and "100"	Desired similarity threshold for the sources, expressed in percentage points. The sources will be assessed for their similarity (e.g., semantic closeness), and the ones that are not close enough to the group will be dropped from the summary.	Setting **simThreshold** to 0 will exclude similarity check and ensure that every source is considered for the summary. Setting **simThreshold** to 100 will likely exclude most sources. Recommended

			settings are: around 30 for a relatively close set of sources, and around 70 for a relatively loose set.
bHtml	string; value can be "1" or "0"	Whether to return result as HTML (1) or not (0)	
bTitle	string; value can be "1" or "0"	Whether to include the title of the summary in the result (1) or not (0)	If the summary is not requested (i.e., **bSummary** equals 0), the value of this parameter is irrelevant.
bSummary	string; value can be "1" or "0"	Whether to include a multidocument text summary of the sources in the result (1) or not (0)	
bConcepts	string; value can be "1" or "0"	Whether to include semantic concepts identified in the sources in the result (1) or not (0)	
groupSentences	string, value can be: "*Group*" or "*Refs*"	Specifies how the sources are referenced in the summary. "*Group*" – sequential sentences from the same source are grouped together in one paragraph, and the source is referenced once after the group. The reference consists of the word "SOURCE:" followed by the HTML hyperlink to the source document. Within the group, the sentences are separated by […]. "*Refs*" – the sentences are grouped the same way as under "*Group*" above, but the reference consists of the source number in square brackets, e.g. [1]. The numbered list of all sources is returned following the summary.	If the summary is not requested (i.e., **bSummary** equals 0), the value of this parameter is irrelevant, as no sources are returned.

Result:

An XML string containing the text elements below. The elements will be present or not depending on the input parameters. The format is oriented at the presentation of the results as HTML page. If you need to get access to individual result elements, e.g. to reformat them, your application will need to parse the string according to the format below.

The string will have HTML formatting if requested by **bHtml** parameter.

Element	Description	Comments

HTML page header	Standard HTML page header	Present if requested by **bHtml** parameter.
List of concepts	Each concept can consist of one word or several words. Each concept is followed by colon and number, corresponding to its semantic weight, e.g. "computer:5". The concepts are separated by a space character. The concepts are returned in no particular order.	Present if requested by **bConcepts** parameter. The maximum number of concepts returned will not exceed 6 times the number of sources.
Separator	 if HTML format is requested, otherwise "\r\n" (CRLF)	
Title	Title of the summary, as specified in the query parameter	Present if requested by **bTitle** parameter.
Separator	 if HTML format is requested, otherwise "\r\n" (CRLF)	
Summary	Multidocument text summary of the sources specified by the allURLs parameter. The summary is prefixed by the word "SUMMARY:". The summary contains references to sources as per groupSentences parameter.	Present if requested by **bSummary** parameter.
Separator	 if HTML format is requested, otherwise "\r\n" (CRLF)	Present if the value of **groupSentences** was set to "*Refs*"
List of sources	List of sources used in the summary, in the following format: - each source is on a separate line - starts with the order number in square brackets, e.g. [1] - is followed by the HTML hyperlink to the source document.	Present if the value of **groupSentences** was set to "*Refs*. The order of sources has nothing to do with their order in the **allURLs** parameter; it is the order in which they are referenced in the summary.
HTML page footer	Standard HTML page footer	Present if requested by **bHtml** parameter.

NOTE: In the examples below, substitute *servername* with the actual server domain you are connecting to.

EXAMPLE 1: get a summary of 2 pages (CNN.com and Reuters.com), in HTML format, with the Title ("news update") and 3-sentences Summary, grouping the sentences as "Group", using HTTP GET:

GET URL:
http://*servername*/svc/contentsvc.asmx/CreateSummary?userName=12345&numSentences=3&numSources=2&query=news%20update&lang=English&simThreshold=30&bHtml=1&bTitle=1&bSummary=1&bConcepts=0&groupSentences=Group&allTitles=CNN%20site%60Reuters%20site&allURLs=http%3A//www.cnn.com%60http%3A//www.reuters.com

Result returned:

```
<?xml version="1.0" encoding="utf-8"?>
<string xmlns="http://sensebot.net/svc/"><html><head><meta http-equiv="content-type"
content="text/html; charset=utf-8"/></head><br><center><b><i>SUMMARY:</i> "news
update"</b></center>

<br><br>10 most popular stories on CNN.com updated every 20 minutes. <b>[...]</b>

<br><br>Last week's angry-mob-with-pitchforks approach to bonuses paid by AIG is giving way
to a more measured approach this week as lawmakers and investors weigh the potential risks of
the proposals before them.
<br>[<i>SOURCE: <a href="http://www.cnn.com">CNN site</a></i>]

<br><br>The United States is fighting a fire in the world economy, but Germany and some other
European countries fear a flood of inflation as a result.
<br>[<i>SOURCE: <a href="http://www.reuters.com">Reuters site</a></i>]

<br><br></html></string>
```

EXAMPLE 2: the same as in Example 1 but without HTML formatting:

GET URL:

http://*servername*/svc/contentsvc.asmx/CreateSummary?userName=1234
5&numSentences=3&numSources=2&query=news%20update&lang=English&si
mThreshold=30&bHtml=0&bTitle=1&bSummary=1&bConcepts=0&groupSenten
ces=Group&allTitles=CNN%20site%60Reuters%20site&allURLs=http%3A//
www.cnn.com%60http%3A//www.reuters.com

Result returned:

```
<?xml version="1.0" encoding="utf-8"?>
<string xmlns="http://sensebot.net/svc/">
SUMMARY: "news update"


10 most popular stories on CNN.com updated every 20 minutes. [...]

Last week's angry-mob-with-pitchforks approach to bonuses paid by AIG is giving way to a more
measured approach this week as lawmakers and investors weigh the potential risks of the
proposals before them.
[SOURCE: <a href="http://www.cnn.com">CNN site</a>]

The United States is fighting a fire in the world economy, but Germany and some other European
countries fear a flood of inflation as a result.
[SOURCE: <a href="http://www.reuters.com">Reuters site</a>]

</string>
```

EXAMPLE 3: the same as in Example 2 but with the sources displayed under the summary (the "Refs" setting):

GET URL:
http://*servername*/svc/contentsvc.asmx/CreateSummary?userName=1234
5&numSentences=3&numSources=2&query=news%20update&lang=English&si
mThreshold=30&bHtml=1&bTitle=1&bSummary=1&bConcepts=0&groupSenten
ces=Refs&allTitles=CNN%20site%60Reuters%20site&allURLs=http%3A//w
ww.cnn.com%60http%3A//www.reuters.com

Result returned:

<?xml version="1.0" encoding="utf-8"?>
<string xmlns="http://sensebot.net/svc/">
SUMMARY: "news update"

10 most popular stories on CNN.com updated every 20 minutes. [...]

Last week's angry-mob-with-pitchforks approach to bonuses paid by AIG is giving way to a more measured approach this week as lawmakers and investors weigh the potential risks of the proposals before them. [1]

The United States is fighting a fire in the world economy, but Germany and some other European countries fear a flood of inflation as a result. [2]

[1] CNN site
[2] Reuters site
</string>

EXAMPLE 4: extract semantic concepts from a set of 2 pages (CNN.com and Reuters.com), in HTML format, using HTTP GET:

GET URL:
http://*servername*/svc/contentsvc.asmx/CreateSummary?userName=1234
5&numSentences=3&numSources=2&query=news%20update&lang=English&si
mThreshold=30&bHtml=1&bTitle=0&bSummary=0&bConcepts=1&groupSenten
ces=Refs&allTitles=CNN%20site%60Reuters%20site&AllURLs=http%3A//w
ww.cnn.com%60http%3A//www.reuters.com

Result returned:

<?xml version="1.0" encoding="utf-8"?>
<string xmlns="http://sensebot.net/svc/"><html><head><meta http-equiv="content-type" content="text/html; charset=utf-8"/></head>BULL:1 CRASH:1 ECONOMY:2 POPULAR:1 PLAN:1 FUNDS:1 AIG:1 GEITHNER BANK PLAN:1 ATTRACT KEY BIPARTISAN:1 TECHNOLOGY:1 HEALTH:2 TREASURY:1

</html></string>