Enabling n-gram data sets to improve grammar quality

With package version 5.6.3, we added the ability to enable n-gram data sets in the grammar engine to detect errors with words that are often confused, like *their* and *there*. The n-gram data sets are available for the group of English dialects, German, French, Spanish, and Dutch languages.

Due to a considerable increase in minimal hardware requirements and a moderate increase in quality, n-grams are not included by default in the standard packages. However, if you still prefer to make use of this option, this guide will help you to setup n-grams for your on-premises installation.

- The n-gram data sets are available for the group of English dialects, German, French, Spanish, and Dutch languages.
- Overall, improvement in the grammar checking accuracy is fairly low (from 0.3% to 1.8%). The result depends a lot on the size of the n-gram data set. The larger the data set, the better the result.
- There is a side effect to having larger data sets. The larger the data, the slower the response time and higher the requirements for hardware (specifically the SSD).

1. Download and unzip n-gram data sets

Contact our support team to get the links to download the latest n-gram data sets for required language(s).

The n-gram data sets are huge and may take from 2.4-14.3GB of SSD depending on the chosen language or their combination. Make sure you have a SSD that fits the space requirements and has at least 10% free space left.

Details about file sizes by languages are represented in the table below.

Language	Zipped size, GB	Unzipped size, GB
English	8.75	14.3
German	1.58	3.06
Spanish	1.68	3.03
French	1.78	3.17
Dutch	1.2	2.4

2. Stop AppServer

It is recommended to stop AppServer before making any changes to the AppServerX.xml file.

3. Specify path to n-gram data sets in AppServer configuration file

- Open the **AppServerX.xml** configuration file for editing.

The default path to the AppServerX.xml file: <WebSpellChecker_Installation_Path>/AppServer/AppServerX.xml

- Find the **PathToNgramData** parameter which is responsible for enabling and configuration of n-gram data sets.

AppServerX.xml

```
<!-- Path to n-gram data sets. Can be used to improve grammar quality. -->
<!-- <PathToNgramData></PathToNgramData>-->
```

- Uncomment the **PathToNgramData** parameter and set a path to unzipped folder of n-grams.

AppServerX.xml

```
<PathToNgramData>your_path_to_ngrams</PathToNgramData>
```

For example, the path for Windows: <PathToNgramData>C:/Program Files/WebSpellChecker/AppServer/NgramData/</PathToNgramData>

The path for Linux: <PathToNgramData>/opt/WSC/AppServer/NgramData/</PathToNgramData>

- Add the **EnableNgramData** parameter inside the **Language** tag for language(s) where n-grams should be enabled.

AppServerX.xml

```
<EnableNgramData>true</EnableNgramData>
```

This is an example of the **EnableNgramData** parameter enabled for American English. You can find the list of language short code (used as Language Id) with the approprialte language in the Default Language guide.

AppServerX.xml

```
<Language Id="en_US">
        <Alias>en</Alias>
        <Alias>am</Alias>
        <GrammarCheckProviderOptions>en-US</GrammarCheckProviderOptions>
        <EnableNgramData>true</EnableNgramData>
        <ThesaurusEnabled>true</ThesaurusEnabled>
                <SpellEngineOptions>
                <Locale>am</Locale>
                <SpellCheckProvider>ssce</SpellCheckProvider>
                <Dictionary FullPath="ssceam2.clx">
                        <ForSuggest>no</ForSuggest>
                </Dictionary>
                <Dictionary FullPath="ssceam2s.clx">
                        <ForSuggest>yes</ForSuggest>
                </Dictionary>
                <Dictionary FullPath="sscema2.clx"/>
                <Dictionary FullPath="keywords.clx"/>
                <Dictionary FullPath="ssceam.tlx"/>
        </SpellEngineOptions>
</Language>
```

4. Start AppServer

As soon as you made the nesessary actions to enable n-gram data sets in AppServerX.xml, start AppServer for the changes to take effect.