

Language Filter Overview

Crimson Hexagon has the ability to filter output by language and can be applied in one of two ways: either using the drop-down menu in the “Setup” tab (preferred) or using a language query in the advanced keywords section. Recently, Crimson Hexagon has updated the way Crimson Hexagon classifies the language of a post. A comparison of the new version and the old version of determining language within the ForSight is summarized below.

Old Version

Formerly, ForSight used metadata received from content source providers in order to determine the language of a post. There are two main reasons why using this metadata was not a reliable way of determining language: certain sources (e.g., Facebook) did not provide metadata and many authors would not change their default language settings when posting in a different language. For these reasons, Crimson Hexagon changed the way Crimson Hexagon classifies language.

New Version

Initiated on March 12, 2013, ForSight now uses its own proprietary algorithm to classify the language of a post, based on probability. Using probabilistic language models, the algorithm is able to classify languages and learn over time (which means it can keep up to date with slang, etc.). The algorithm simply finds the language with the highest probability, and uses that one. If there are multiple languages with high probability (i.e., Norwegian and Swedish are very similar), the algorithm will then rely on metadata to try to find the correct language.

Note: If a monitor has a date range that runs through March 12, 2013, expect to see a significant shift in volume. Since this is the date Crimson Hexagon released its new language-classifying algorithm, volume shifts in the data are expected.

What Method to Use

Similar to using a location filter, it is highly recommended that you use the drop-down menu bar on the “Setup” tab whenever possible. The new algorithm fully supports 16 different languages, which are located at the top of the drop-down menu: Arabic, Chinese, Dutch, English, French, German, Indonesian, Italian, Japanese, Korean, Malay, Portuguese, Russian, Spanish, Swedish and Turkish. After thorough testing, Crimson Hexagon can assure that these languages will provide accurate results without showing any strangeness in the visualizations or metrics of the monitor.



If you would like to run an analysis on a language that is not currently one of these 16 "supported" languages, you can still select any of the other 24 languages from the lower portion of the drop down. Please note that Crimson Hexagon cannot assure the reliability of these other languages, and there is a possibility of some disproportionate visualizations and metrics.

It is still possible to use the Language field specifier when working with Advanced Keywords, but we recommend using the drop down. If you choose to include the Language field specifier in your keywords, here is how you would do that:

1. Find the [2-letter language code](#) of the desired language for analysis
2. Enter language:en for the English language or language:hr for Croatian in the advanced keywords section when building a monitor.

Here is a list of the 40 languages available in Crimson Hexagon, preceded by the 2-letter language code:

ar - Arabic

az - Azerbaijani

bg - Bulgarian

ca - Catalan; Valencian

cs - Czech

da - Danish

de - German*

el - Greek, Modern

en - English*

es - Spanish; Castilian*

eu - Basque

fa - Persian

fi - Finnish

fr - French*

he - Hebrew, Modern>

hi - Hindi

hr - Croatian

hu - Hungarian

id - Indonesian*

it - Italian*

ja - Japanese*

kk - Kazakh

ko - Korean*

ku - Kurdish

lt - Lithuanian

ms - Malay*

nl - Dutch*

no - Norwegian

pl - Polish

pt - Portuguese*

ro - Romanian, Moldavian (Romanian from Republic of Moldova)

ru - Russian*

sk - Slovak

sl - Slovene

sr - Serbian

sv - Swedish*

th - Thai



tr - Turkish*

ur - Urdu

zh - Chinese*

*Indicates a language that is supported by our proprietary language algorithm.